SYSTEMATICS OF CETACEANS USING RESTRICTION SITE MAPPING OF MITOCHONDRIAL DNA

DEREK PAUL OHLAND

Thesis submitted in fulfilment of the requirements for the degree of
Master of Science (Med.)
in the Department of Chemical Pathology in the Faculty of Medicine at the
University of Cape Town

March 1992

The copyright of this thesis vests in the author. No quotation from it or information derived from it is to be published without full acknowledgement of the source. The thesis is to be used for private study or non-commercial research purposes only.

Published by the University of Cape Town (UCT) in terms of the non-exclusive license granted to UCT by the author.

University of Cape Town

I, **DEREK PAUL OHLAND**, hereby declare that the work on which this thesis is based is original (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or in any other University.

I empower the University to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

Signed by candidate

Signature removed

25 March 1992

A burning Tripod tells thee thou hast found
The deepest art below the deepest ground;
And by its light the Mothers thou wilt see -
Some sit, and others stand, or, it may be,
In movement are. Formation, Transformation,
Eternal Play of the Eternal Mind,
With Semblances of all things in creation,
For ever and ever sweeping round.

Johann Wolfgang von Goethe (1749 - 1832)
Faust Part II
Translated by John Anster

ACKNOWLEDGMENTS

I wish to express my gratitude to :-

Prof. Eric Harley, for his friendship, initiation and interest in this project.

Fadial Essop, for his patient tutoring.

Dr. Peter Best, for his kind assistance and the donation of otherwise unobtainable biological material.

Prof. M A Cluver for study leave.

Mrs M Compagno-Roeleveld, for her consideration during my writing up.

Everybody in the Lab. for the congenial working atmosphere.

My parents, for encouraging freedom of thought.

Lucy, for so much (including the word-processing of this thesis).

ABSTRACT

A phylogenetic study of eleven cetaceans was undertaken using Restriction Endonuclease Maps (RSM) of mitochondrial DNA (mtDNA). One species from the suborder mysticeti (baleen whales) was sampled, and of the ten odontocetes (toothed whales) sampled two were from the family Ziphiidae (beaked whales) and eight were from the family Delphinidae (dolphins) (each representing a different genus).

The primarily opportunistically obtained (i.e. from strandings or accidental death in commercial trawl nets) heart tissue generally yielded high quantities of mtDNA which is needed for double digest fragment analysis. The mtDNA extracted from the sampled taxa was cleaved with fifteen different six-base Restriction Enzymes (RE's). Using the three-way method of analysis and aided by the computer program Resolve (Ver. 2.7) (Harley, unpublished), RSM's were constructed. Distance (Neighbor-Joining and Fitsch-Margoliash) and cladistic (Maximum Parsimony and Bootstrap) methods were used to infer phylogenies. The baleen whale was used as an outgroup for the cladistic analysis.

Both the distance and both the cladistic methods produced the same single topology, which is concordant with morphologically based classifications. The two differences (within the Delphinidae), viz. *Grampus*' most basally rooted position and *Cephalorhynchus*' grouping with the Delphininae are of taxa whose groupings are unresolved in the morphologically based classifications.

Using Brown *et al*'s (1979) molecular clock, very recent divergence times at the generic, family and suborder levels were obtained, when compared to fossil based estimates. Using the odontoceti/mysticeti split the base substitution rate of cetacean mtDNA was estimated to be much slower than that of terrestrial mammals (0,3% compared to 1,0% Myr $^{-1}$). A similarly slow rate was calculated for cetacean nuclear DNA (nDNA) (0,09% Myr $^{-1}$) (Schlötterer *et al*, 1991). It remains an unresolved issue as to whether the base substitution rate of cetacean DNA is slower than terrestrial mammals or whether the fossil evidence needs to be reinterpreted. The time of the mysticeti/odontoceti split is palaeontologically uncertain and the suggested monophyletic status of the extant suborders has been questioned, thus making the calculation of cetacean base substitution rate risky. Equally, the incomplete fossil record can lend itself to misinterpretation.

LIST OF ABBREVIATIONS

1. NUCLEIC ACID TERMS

dATP(A)	Deoxyadenosine triphosphate
dCTP(C)	Deoxycytidine triphosphate
dGTP(G)	Deoxyguanosine triphosphate
dTTP(T)	Deoxythymidine triphosphate
^{32}P-dCTP	Cytidine triphosphate radioactively labelled with Phosphorus-32 (a position)
kb	kilobase pair
bp	base pair
DNA	Deoxyribonucleic acid
mtDNA	Mitochondrial deoxyribonucleic acid
nDNA	Nuclear deoxyribonucleic acid
scnDNA	Single copy nuclear DNA
DNAse	Deoxyribonuclease
D-loop	Displacement loop
PCR	Polymerase chain reaction

2. UNITS

k, m, μ, n,	kilo-, milli-, micro-, nano- (prefixes)
m, l, g	metre, litre, gram
Ci	Curie
°C	Degrees Celcius
mol	Moles
M	Molar
N	Normal
g	Centrifugal force
V	volt

3. CHEMICALS

EDTA	Ethylenediaminetetraacetic acid
EtBr	Ethidium Bromide
EtOH	Ethanol
SDS	Sodium Dodecyl Sulphate
SSC	Sodium Chloride - Sodium Citrate buffer
Tris	Tris (hydroxymethyl) amino methane
TE	Tris EDTA buffer
STE	Saline Tris EDTA buffer
HCl	Hydrochloric acid
H_2O	Water
NaOH	Sodium hydroxide
KGB	Potassium Glutamate buffer
CsCl	Cesium Chloride
TAE	Tris Acetate EDTA buffer
NaPP	Sodium Pyrophosphate

CETACEANS SAMPLED

Caperea marginata	Pygmy right whale
Mesoplodon layardii	Layard's beaked whale
Hyperoodon planifrons	Southern bottlenose whale
Globicephala melas	Long-finned pilot whale
Feresa attenuata	Pygmy killer whale
Grampus griseus	Risso's dolphin
Lagenorhynchus obscurus	Dusky dolphin
Cephalorhynchus heavisidii	Heaviside's dolphin
Delphinus delphis	Common dolphin
Tursiops truncatus	Bottlenose dolphin
Stenella coeruleoalba	Striped dolphin

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure 1 : A group of Heaviside's dolphins cavorting off the West Coast of Southern Africa
Photograph: M Mittlemeyer

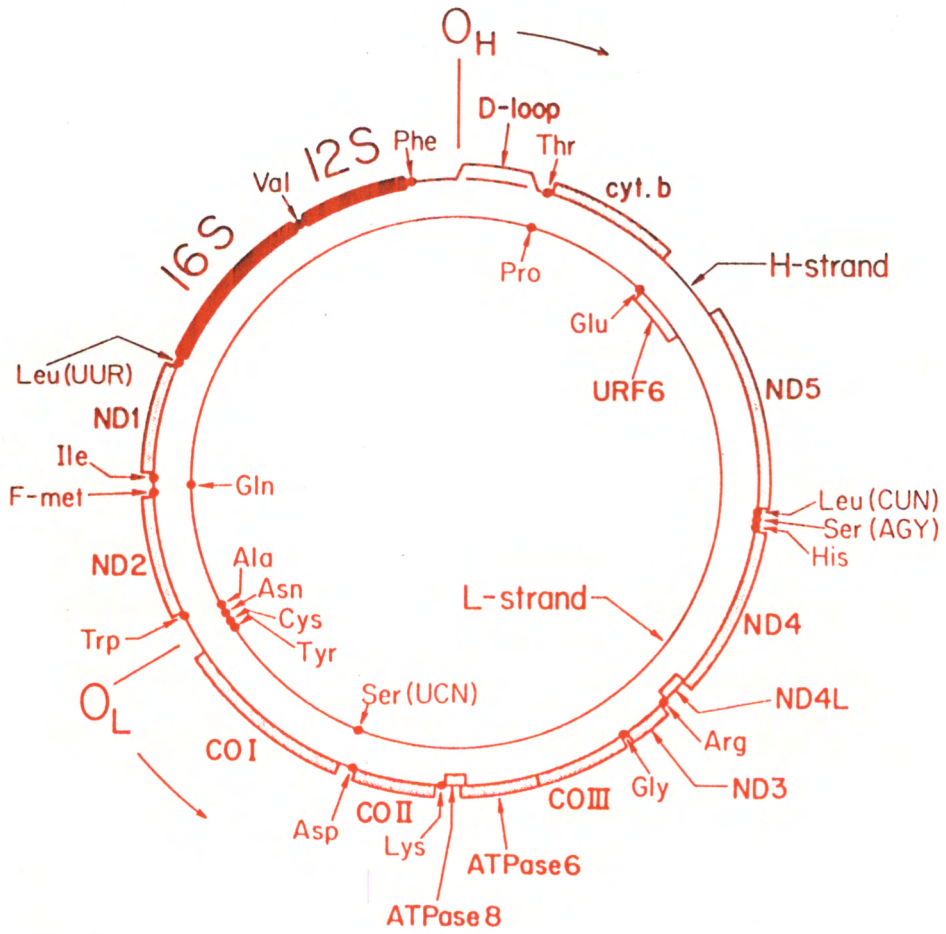


Figure 2 : Map of the 16 500 base pair mammalian mitochondrial genome. The map is split into two parts, the outer circle indicating those genes transcribed from the H (heavy)-strand DNA and the inner circle indicating those genes transcribed from the L (light)-strand DNA. RNA genes (for 16S and 12S rRNA or for tRNA's) are indicated in black, whereas protein-coding genes are coloured. Note that whenever a space exists on one strand, it is filled by a gene on the other strand. In the designation of codons recognised by certain tRNA's, Y stands for pyrimidine, R stands for purine, and N stands for any base. Adenosine triphosphate (ATP)-ase 6 and 8 are components of the mitochondrial ATPase complex. CO I, CO II and CO III are cytochrome oxidase subunits. Cyt *b* stands for cytochrome *b*. ND1-5 code for components of the respiratory chain NADH dehydrogenase and were previously designated URF1-5 (for *unidentified* open *reading frame*). URF6 remains an open reading frame of unknown function. O_H and O_L designate the origins of DNA replication for the H (heavy) strand (which is initiated first) and for the L (light) strand respectively. (Taken from Watson *et al*, 1987.)

CHAPTER 1

INTRODUCTION

1.1 GENERAL CONCEPTS OF PHYLOGENETICS

The most basic element in biological systematics is the individual organism. Even at this elemental level the process of change is observable. Systematic studies do not deal with constant units, but rather with dynamic living forms (Hennig, 1966, p.6). Torrey summarises :- "each describable single form is only an arbitrary portion of the whole that is determined by the point in time chosen" (ref. taken from Hennig, 1966, p.6.).

Hennig (1966) further states that an individual, because of its predisposition to vary, will occupy different positions in most systems during the course of its life, and that therefore the "ultimate" element of the biological system is what he terms the "character-bearing semaphorant" which he defines as "the individual during a certain, however brief, period of time" (Hennig, 1966, p.6).

The holomorphic (or definitive set of characters) that define the semaphorant are its physiological, morphological and psychological (ethological) properties.

From the above it is clear that Hennig (for the purposes of systematics) recognises the state of flux of "the life forms", be it the semaphorant, individual organism or species, and that any qualitative analysis thereof must be made with reference to a time frame. Succinctly stated, Hennig recognises the transformative nature of the holomorphy of the semaphorant.

Phylogenetics (defined as "the sequence of events involved in the evolution of a species") (Collins, 1979) rests on the basic assumption that there is orderliness in the process of change and that this order consists of an hierarchy of patterns of similarity (Eldridge *et al*, 1980)

The theory of evolution itself, defined by Collins (1979) as "a gradual change in the characteristics of a population of animals or plants over successive generations (that) accounts for the origin of existing species from ancestors unlike them", thus equally assumes the orderly nature of change. Evolution implies ancestry and "descent with modification" and that this modification can be described by measuring the transformation of intrinsic properties (Eldridge *et al*, 1980). The aim of phylogenetics is to reconstruct the evolutionary history of life forms (taxa) as closely as possible to the unknown reality, by using various empirically based deductive techniques (Li *et al*, 1990).

Given the ability to enter into a "supra-geological" dimension of time in which evolutionary events that usually span millions of years could be telescoped into a few weeks or so, we would be endowed with a "bird's-eye" view of the origins of species. Probably the most fascinating thing that would strike us (apart from the interdependent nature of all things) would be the extremely malleable or transformative nature of life.

In a constant interaction between genotype and environmental pressure phenotypes would be seen to almost fluidly adopt advantageous dispositions and selectively evolve certain existing characteristics, whilst disregarding obsolete adaptive systems, so that the optimal mode of functioning within genetic and environmental limitation is attained.

Phylogenetics would be made simple. A detailed set of relationships between taxa within any order or between orders themselves would be immediately observable. Evolutionary milestones, points of radiation and ancestral histories could be empirically obtainable.

If we stretch our imagination still further and increase the speed of time to maximum, there would most likely be a debate amongst the phylogeneticists in the time machine as to whether the concept "species" is actually justifiable, as what would be observed during this extreme telescoping of events would be constant change; a formation and transformation of forms so fluid and continuous that the only constant to be observed would be the process of change itself.

Having no such Jules Verne contraption at our disposal, the minds of some men such as Darwin (1859) soared instead to the great heights of creating the theory of evolution.

Essentially, by measuring the transformative nature of a species' distinguishing characteristics over a period of time the tentative reconstruction of the evolutionary histories and the classification of groups of taxa is made possible (Darwin, 1859) (Eldridge *et al*, 1980).

1.2 A BIOLOGICAL CONCEPT OF HOMOLOGY

In general, homology means similarities derived from inferred common ancestry. Re-phrased, the degree of similarity (between species) is used as a measure to construct (their) phylogenies (Eldridge *et al*, 1980). Darwin's "On the Origin of the Species" is based on such a concept (Eldridge *et al*, 1980). As will be further discussed under Chapter 7, homoplasies are sources of error variance that can confound phylogenetic reconstructions based on such shared similarities.

In the words of Darwin himself (1859, p.411), "All organic beings are found to resemble each other in descending groups." And again, in describing his diagram of the theory of evolution, he continues -

> " .... and he will see that the inevitable result is that the modified descendants proceeding from one progenitor become broken up into groups subordinate to groups. So that we here have many species descended from a single progenitor grouped into genera; and the genera are included in, or sub-ordinate to, subfamilies, families, and orders, all united into one class. Thus the grand fact in natural history of the subordination of group under group ...." (pp.412-413).

1.3 MOLECULAR APPROACH TO PHYLOGENETICS

There are many variables that contribute towards the evolution of a species. Bio-geographical factors, climatic changes, adaption to ecological niches, (Fordyce, 1980), population pressure on resources, altered food-chain hierarchies, altered ecosystems (Gaskin, 1982) and between species competition are some examples. Intimate symbiotic relationships between organelles such as eukaryotes (cells) and bacteria have greatly affected the evolutionary history of multi-cellular organisms (Darnell *et al*, 1986).

Genetic potential and gene mutations are other obvious major contributing factors (Darnell *et al*, 1986). It is not within the scope of this thesis to elaborate on the possible influences that contribute towards the evolution or obliteration of a species. Rather it is an attempt to -

(a) re-construct the phylogeny at the generic level of eight taxa from family Delphinidae; and

(b) estimate the time of divergence from the last common ancestor, at the generic, familial and sub-order levels of Cetacea.

In the past, most phylogenetic studies made use of morphological, palaentological and behavioural based characteristics as a qualitative method to determine the degree of relatedness between species and their evolutionary histories (Eldridge *et al*, 1980).

Recently, however, a number of molecular techniques have been evolved which can be used for phylogenetic studies (Hillis *et al*, 1990). Such molecular methodologies enable a data base to be obtained through the sampling or characterisation of macro-molecules such as proteins, nuclear DNA (nDNA) or organelle DNA (i.e. mitochondrial and chloroplast DNA).

1.3.1 A Molecular Concept of Homology

The degree of similarity between DNA or protein molecules can be empirically observed and quantitatively measured. Specifically, homologous characters are used to infer the degree of similarity (or relatedness).

This inferred degree of homology need not stem solely from common ancestry as two other causes can contribute towards the degree of (homologous) similarity between the molecules. Hillis *et al* (1990) distinguish between three types of homologies only one of which, the orthologous sequence, can be used to infer phylogeny of species. Orthologous homologous sequences attain their similarity through common ancestry. Paralogous homologies are those sequences that diverged after gene duplication. Homologies that arise via lateral gene transfer (for example through retroviruses) are called xenologous sequences.

The inability to differentiate between paralogous and orthologous sequences can result in the reconstruction of the correct gene phylogeny but not in the correct species phylogeny.

1.3.2 The ideal deoxyribonucleic acid molecule (DNA) data set would be a complete description of the genome itself. However, the sheer enormity of the DNA molecule (nDNA has approximately 3 billion base pairs in man for example, whereas mtDNA has approximately 16 400) (Watson *et al*, 1987) makes this a virtually impossible task, considering the limitations of current techniques (Hillis *et al*, 1990).

It is interesting to note that undaunted researchers have indeed already sequenced the complete mitochondrial genome for a number of animals, and that the Herculean task of sequencing the entire Nuclear DNA (nDNA) genome of the human has been undertaken, a project headed by Watson, one of the two Nobel prizewinners who formulated the complementary double-helix structure of the DNA molecule in the mid-fifties (Watson *et al*, 1987).

The present study samples the mtDNA genome by using the endonuclease restriction mapping technique. This method, along with other molecular techniques, will be described and discussed in the following pages. The data sets thus obtained are subsequently analysed using both distance and cladistic statistical approaches. These results are then compared with each other, with morphologically-based systematic studies, and with the findings from other molecular approaches. Datings of points of radiation obtained from distance data analysis are compared and discussed with other studies whose findings are based on nDNA analysis, allozyme studies and of course with palaentological research deductions.

Possible scenarios of the evolutionary histories of the cetaceans under study are created and discussed in the light of the above research.

1.4 "MOLECULES VERSUS MORPHOLOGY" CONTROVERSY

Rather than enter into the "molecules versus morphology" controversy, it is suggested that the incorporation of both molecular and morphological data will provide a more comprehensive understanding of phylogenies than either one of the approaches can do in isolation.

To quote Hillis *et al* (1990, p.4) -

> " ... the real concerns for the practising systematist are whether the characters examined exhibit variations appropriate to the question(s) posed, whether the characters have a clear and independent genetic basis, and whether the data are collected and analysed in such a way that is is possible to compare and combine phylogenetic hypotheses derived from them."

1.5 INTENDED AREAS OF EXPANSION ON THE PRESENT STUDY

1.5.1 It is the author's intention to compile a more comprehensive phylogeny of the Odontoceti using the restriction endonuclease mapping technique, by including (mostly as yet unobtained) new species.

1.5.2 It would be interesting to compare inferred phylogenies using both restriction site mapping and sequencing techniques.

1.5.3 Finally, the success of the Polymerase Chain Reaction (PCR) technique (Hillis *et al*, 1990) in amplifying DNA from old material will make it an indispensable technique in the utilisation of the fragments of old flesh that are still attached to the skeletal material that is housed in the comparative osteology collections. The bone itself can be used as a source of DNA. The PCR technique will enable phylogenetic studies to be undertaken on rare or inaccessible cetaceans whose skeletal material has slowly been collected over the decades.

It is hoped that the following dissertation will demonstrate the validity of molecular techniques as being constructive to phylogenetic research. It is also hoped that the results thus far obtained will contribute towards a greater understanding of cetacean systematics and their evolutionary history.

CHAPTER 2

EVOLUTIONARY HISTORY OF CETACEA

2.1 ANCESTRAL LINEAGE OF THE ORDER CETACEA

The family Mesonychidae of the order Condylarthra have long been considered ancestral to the modern artiodactyls (the cloven-hoofed ungulates). Comparative skull morphology and biochemical comparative analysis of fetal blood sugar, blood composition, chromosomes, insulin, uterine morphology and tooth enamel micro-structure between modern mammals suggest that cetaceans are most like artiodactyls and hence probably share the mesonychidae as a common ancestor with the extant ungulates (Barnes, 1984). Albumin comparisons indicate that cetaceans are most closely related to the ruminant artiodactyls (Lowenstein, 1985). Equally, van Valen (1966) and Szalay (1969) concluded from fossil evidence that archaeocetes (a primitive form of toothed whale) evolved from mesonychid condylanths.

The now extinct Archaeoceti (Eocene period - 38 to 55 million years [Myr] ago) were the most ancient group of cetaceans known (Barnes, 1984, March). Whether or not they are directly ancestral to the later whales which appeared from the mid-Oligocene on (approximately 28 Myr ago to the present) is debatable (Barnes, 1984).

The order Cetacea (from the Greek ketos or Latin cetus, meaning large sea animal) is classically divided into three major groups.

These are the previously mentioned ancient whale group, the Archaeoceti, and the two modern suborders, the odontocetes (from the Greek odontos meaning toothed) and the mysticeti (from the Greek mystax, meaning moustache, which refers to the baleen in the upper jaw) (Barnes, 1984, March).

2.2 STATUS OF CETACEA: MONOPHYLETIC OR POLYPHYLETIC ?

It is a controversial issue whether the archaeoceti, mysticeti and odontoceti constitute a monophyletic (sharing a common ancestor) or parayphyletic (arising from separate origins) group. The second issue is whether the odontoceti and mysticeti stem from the same or different lineages. This latter issue is discussed in some detail, as the mysticeti/odontoceti split is used to calculate the base substitution rate of cetacean mtDNA.

2.2.1 In support of the monophyletic status of cetacea, Barnes (1984, pp. 139-140) states that -

> " All cetaceans have the following unique suite of shared derived characters (synapomorphies) :- (1) loss of anterior palatine foramina; (2) large falcate processes of the basioccipital; (3) peribullary and pterygoid air sinuses present as diverticular from the middle ear sinus; (4) tympanic bulla involuted and inflated (5) supra-orbital process of the frontal large, horizontal and tabular; (6)

hypoglossal foramen in the basioccipital located either at the apex of or inside the jugular notch; (7) large mandibular foramen; and (8) scapula with the supra spinatus fossa reduced and with acromion and coracoid processes parallel and directed anteriorly."

Similarly in support of the monophyletic status of the two existing suborders of cetacea Barnes (1984) lists eight synapomorphies that are common between odontoceti and mysticeti, but which are not found in the extinct Archaeocetes. Analyses of whale chromosomes (Gaskin, 1982) also indicate a common ancestry for baleen and toothed whales. Whitmore and Sanders' (1976) conclusion that all known odontocetes and mysticetes date only as far as the Oligocene (25 to 28 Myr ago) chronologically supports the two suborders' singular ancestral lineage with Archaeoceti. Barnes (1984) substantiates this by demonstrating the "morphologically intermediate" nature of the Oligocene toothed and baleen whales when compared with the Archaeoceti (Eocene) and with the post-Oligocene (Miocene and later) suborders. That the mysticeti evolved from a toothed whale is also evidenced by the presence of embryological dentition before the development of baleen plates. Similarly, some primitive mysticetes had functional teeth as well as developed baleen (Gaskin, 1982).

2.2.2 Questionable Monophyletic Status of the Extant Suborder of Cetacea

Yablokov *et al* (1972) are not so certain that the origin of baleen and toothed whales is monophyletic.

These authors maintain that the continual debate between proponents of the two concepts of the development of cetaceans is essentially fruitless, as in the absence of fossil records tentative evidence can be found to support either one of the hypotheses, which are :-

(a) That the Cetacean groups originate from different ancestors and that their common features are the result of convergent evolution (polyphyletic); and

(b) That the similarities between the two suborders are shared derived characters and are thus indicative of a common ancestor (monophyletic).

Yablokov *et al* (1972) list thirty-seven skeletal features of similarity in baleen and toothed whales, but continue to say that "almost all" of these features are also typical of the Sirenia order and in part for Pinnipedia, thus these data cannot be used to support any definitive phylogenetic conclusions. As Yablokov *et al* (1972, p.432) themselves state -

> " Thus, most of the features the aggregate of which characterise the order of cetacea in its contemporary range, are also encountered in other mammalian orders, so that the presence of these traits in odontoceti and mysticeti cannot serve as irrefutable proof of the common origin of these groups."

In a case against the monophyletic status of the two extant suborders of cetacea, Yablokov *et al* find it anomalous that "fully formed baleen whales" (p.432) (Family: Cetotheriidae) suddenly appear in the mid-Oligocene (35 - 40 Myr ago). The authors maintain that the development of a set of features that characterise a new mammalian genus (let alone family) takes "several (usually many dozen) million years" (p.433), a view which tends to support the possible paraphyletic nature of odontoceti and mysticeti.

2.3 RADIATIONS OF CETACEA

Fossil evidence indicates three major periods of radiation of the cetacean order :-

2.3.1 The first is the evolution of the Archaeoceti from their proposed ancestors the Mesonychidae during the Eocene period (38 to 55 Myr ago).

2.3.2 A palaentologically blank early Oligocene period is followed by evidence of a second radiation with the emergence of the baleen and toothed whale suborders from their proposed common ancestor, the Archaeoceti, during the late Oligocene - early Miocene period (approximately 20 to 30 Myr ago).

2.3.3 The period during which forms of modern Cetacea (both baleen and toothed whales) became most abundant constitutes the third major radiation and occured from the mid-Miocene to the present (17 Myr ago +) (Barnes *et al*, 1985).

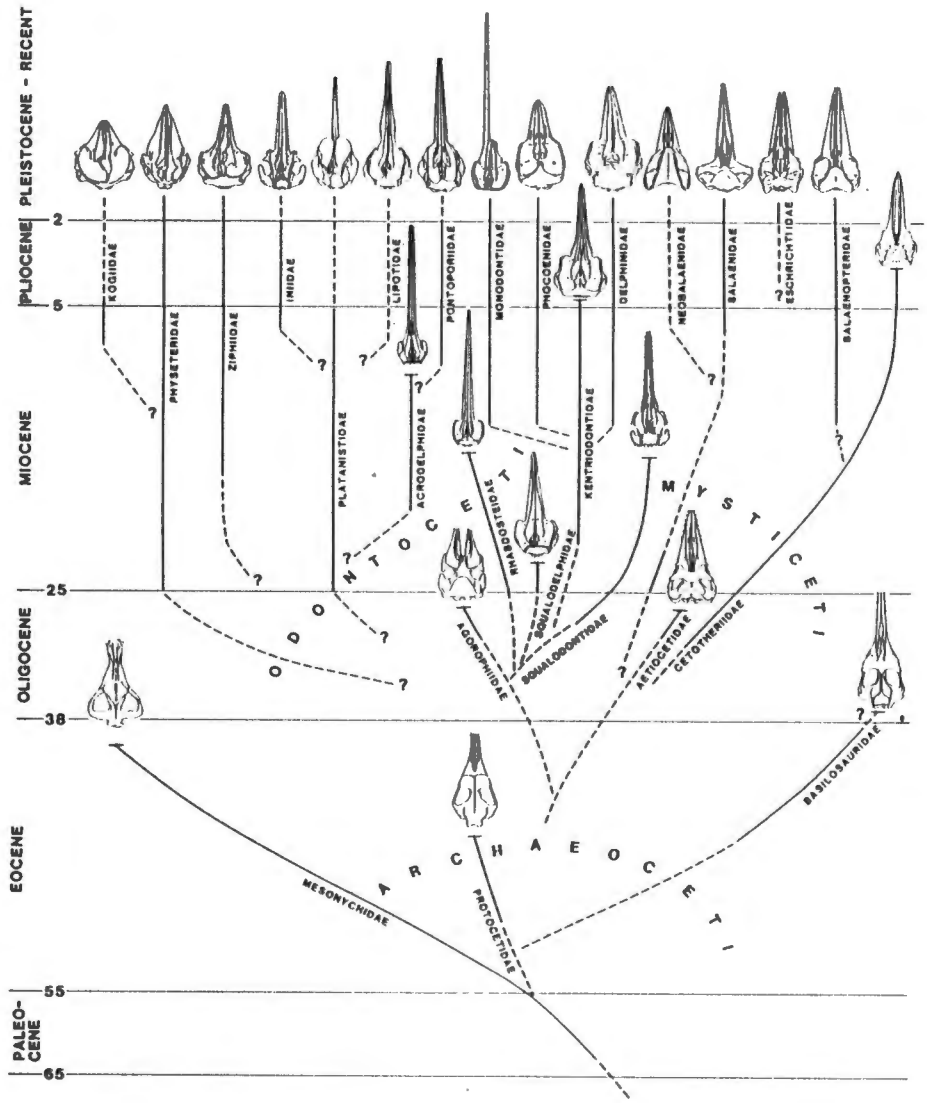


Figure 3: Phylogony of the cetacean families (taken from Barnes *et al*, 1985).
Dotted lines indicate no fossil record. In most cases skulls are of the type genus of the family. Numbers at epochal boundaries are millions of years.

2.4 LIMITATIONS OF PALAEONTOLOGICAL EVIDENCE

As can be immediately ascertained from Barnes' phylogeny of the cetacean families (Fig. 3), there are major palaeontological gaps, particularly in the direct ancestral lineages and the relationships between the extant families. In fact, none of the existing families are palaeontologically directly linked with each other, therefore ancestral histories cannot be unequivocably ascertained. This is a frustrating palaeontological fact, and the lack or sparsity of fossil evidence invariably leads to the creation of tentative rather than immutable phylogenies.

As Yablokov *et al* (1972, p.429) state, after having discussed that skeletal attributes shared by various extinct cetacean families can be attributed with "equal justification" to either convergence or to true phylogenetic relationships -

> " The example discussed above with evaluation of correlations between extinct forms indicates that the widespread conviction of neontologists that palaeontological methods are omnipotent with respect to reconstruction of phylogenesis is fallacious: at the most important state of investigation, the palaeontologist, like the contemporary zoologist, has to derive a conclusion as to relationship merely on the basis of similarity"

Schlötterer *et al* (1991, p.65), too, succintly state -

> " .... the oldest fossils that are clear predecessors of the odontocete and mysticete whales come from the early Miocene and might be as young as 20 million years old. Older fossils cannot be unequivocally ascribed to direct ancestors and could therefore represent extinct parallel lineages."

2.5 FACTORS CONTRIBUTING TO THE LACK OF DETAILED PHYLOGENETIC RELATIONSHIP AMONGST THE CETACEA

Two main factors have contributed to the lack of an accurate description of the phylogenetic relationships amongst the cetacea, these being -

(a) the sparsity of fossil records; and

(b) the remote or inaccessible habitat of cetaceans.

Many species are pelagic and all spend a substantial part of their existence under water. Their constant movement increases the difficulty of controlled observation and collection of specimens. Similarly their pelagic existence reduces the availability of fossils since the majority of animals die in the open seas, resulting in the gradual fragmentation of their bodies. Only occasional strandings provide for keeping the skeleton intact and even then the chances of the skeleton being covered by sedimentary deposits conducive to good fossilisation are rare (Schafer, 1972).

Barnes *et al* (1985) state that the comprehension of marine mammal history is also biased by factors such as the presumption that all oceans and some fresh-water systems supported marine mammals through the Cenozoic era, but that fossils have not been found in all these areas primarily due to the lack of the formation of fossil-preserving sedimentary rock. The retrieval of fossils is also dependent upon natural erosion, man-made excavations, financial support, and simply 'being in the right place at the right time'.

In an excellent summation of the importance of the quality of fossils in the understanding of the evolutionary history and classification of any group, Barnes *et al* (1985, p.16) summarises :-

> " What we know of the fossil history of any group influences our understanding of the classification, evolution and ecology of its modern representatives. Inherent in palaeontologic analysis is the study of bones and teeth, and all the subsequent interpretations *hinge on the quality of the fossils* and of the base line descriptive work" (my italics).

2.6 ORIGINS OF PERTINENT LINEAGES WITHIN CETACEA

2.6.1 Delphinidae

The present study is primarily concerned with the phylogenetics of the dolphin group. The family Delphinidae is part of the super-family Delphinoidea, which also includes the Phocoenidae (porpoises), Monodontidae (Belugas and Narwhals) and the extinct Kentriodontidae.

All three living families probably evolved from the Kentriodontidae, although there is no direct fossil evidence to support this (Barnes *et al*, 1985). The earliest fossils from the modern delphinid family are from the late Miocene period (Gaskin, 1982). Several other groups of dolphin-like families existed in the earlier Miocene (23 - 25 Myr ago), but according to Barnes (1984) they do not strictly possess the characteristics of the modern delphinids, and are thus considered to be extinct forms of parallel lineages.

Trofimov and Gromova (1962) (Taken from Yablokov *et al*, 1972) extend the modern Delphinidae lineage right back to the start of the Miocene (25 Myr ago) and Fordyce (1980) estimates their time of origin at mid-Miocene (approximately 15 Myr ago). "Modern" delphinids became most abundant during the Pliocene period (2 - 5 Myr ago) (Gaskin, 1982), although specimens resembling genera of modern Delphinidae, such as *Stenella* and *Tursiops*, have been discovered in very late Miocene deposits (5 - 8 Myr ago) (Barnes, 1976).

2.6.2 Ziphiidae

Fossil evidence of the beaked whales (family Ziphiidae) extends continuously back to the middle Miocene (approximately 15 Myr ago), (Barnes *et al*, 1985), giving them the second longest fossil lineage amongst living families. The sperm whale's (family Physeteridae) lineage is palaeontologically the longest known, and extends back to the very early Miocene (approximately 25 Myr ago) (Fordyce, 1980).

In an interesting paper Mead (1975) describes a beaked whale's rostrum which was found inexplicably in freshwater deposits from the mid to late Miocene epoch (approximately 10 Myr ago). From this he concludes that the living Mesoplodon and possibly Hyperoodon evolved from the advanced ziphiid *Belemnoziphius* or *Proroziphius*, rather than from a more primitive form.

Heyning (1989) undertook a cladistic analysis at the super-familial level of the extant odontoceti based on facial anatomy and other morphological data. His results indicate the evolution of extant odontocetes in the following order; the physeterids first, followed by the ziphiids, platanistids, iniids and delphinoids.

2.6.3 Neobalaenidae

In order to obtain a molecular-based dating estimate of the time of radiation between the odontoceti and mysticeti an opportunistic fresh stranding of *Caperea marginata* (pygmy right whale) from the family Neobalaenidae was sampled. Although the mysticeti have a broken fossil record extending right back to the Oligocene, there is only one reputed fossil member of this family (*Caperea simpsoni* from South America), which makes this unique baleen whale's history an enigma (Barnes, 1984).

2.7 PERTINENT ISSUES RELATIVE TO THE PRESENT STUDY ON THE EVOLUTIONARY HISTORY OF CETACEA

The most pertinent issues relative to the present study on the evolutionary history of cetacea are :-

(a) The age of modern cetacean groups;

(b) During which periods the major radiations took place; and

(c) Whether the early fossil groups are directly ancestral to the extant families, or whether they represent extinct parallel lineages.

2.8 SPECIFIC CONCERNS OF THIS STUDY

Specifically, this study is concerned with -

2.8.1 The evolutionary history at the generic level within the family Delphinidae, including the reconstruction of ancestral lineages and the estimation of the relative (and perhaps absolute) timing of radiation events.

2.8.2 At the family level, the degree of relatedness and time of radiation of the extant families Ziphiidae and Delphinidae.

2.8.3 Similarly at the suborder level, to determine the degree of genetic relatedness between the odontoceti and mysticeti.

CHAPTER 3

CLASSIFICATION OF EXTANT CETACEANS

3.1. ANATOMICAL ADAPTIONS OF MODERN CETACEANS

Living cetaceans share basic anatomical and physiological characteristics with other mammals, but still demonstrate a high morphological and anatomical adaption to their adopted marine environment. Anatomical adaptions include a hypodermal blubber layer for food storage and thermo-regulatory control, developed kidneys for altered saline balance, high titre of muscle globin for rapid transfer of oxygen to the cellular level and an exceptional resistance to lactic acid accumulation. Morphological specialisation includes a streamlined body, the development of flukes as a means of propulsion, the loss of tetrapod characteristics, the loss of most of the pelvic girdle, loss of hair, dorsally placed nostrils, no external ears, and the development of a dorsal fin for thermoregulation and hydro-dynamic control, the last feature being most prevalent in the smaller toothed whales (Gaskin, 1982).

3.2 DIFFERENTIATION OF SUB-ORDERS WITHIN EXTANT CETACEANS

The odontoceti and mysticeti suborders are separated primarily on the basis of feeding strategies and accompanying morphological transformations (Fordyce, 1980).

3.2.1 The mysticeti are filter feeders, devouring large amounts of zooplankton and small fish, and as such have developed the highly specialised baleen plates in place of teeth. Baleen is formed from keratin and is epidermal in origin. Baleen plates are fibrous structures that grow in rows from the maxilla, similar to teeth on a comb. Using its loosely-articulated jaws and expandable mouth cavity a balaenopterid whale gulps a volume of water and with a powerful tongue expels it through the baleen plates which entrap the organelles upon which it feeds.

3.2.2 Odontoceti, as did the more primitive Archaeoceti, feed upon a discrete prey such as squid and fish. To this end the odontoceti have evolved a highly-developed sense of echo-location, which the Archaeoceti did not possess (Fordyce, 1980). A series of high-frequency clicks is produced in the nasal passages, focussed through the melon on the face (which apparently acts as an acoustic lens), and is then projected into the environment. Reflected sound waves are transmitted via the lower jaw to the ear region, which is specially developed to allow directional hearing (Norris, taken from Barnes, 1984). Odontocetes have also developed their teeth in various ways to facilitate their hunting abilities as well as for social interactions (especially males only in the ziphiids). For example, some species have increased the number of their teeth (as have most dolphins), others have evolved more simple tooth structures such as single roots and conical crowns (all odontocetes except ziphiids), whereas some beaked whales have developed protruding tusks (Barnes 1984).

3.3 HIGHER LEVEL CLASSIFICATION OF THE EXTANT CETACEAN

The mysticeti are divided into four families that include a total of eleven taxa.

The odontoceti form a much larger group, which is divided into nine families consisting of some sixty-seven taxa. Some odontoceti families such as Platanistidae (river dolphins) Iniidae (Amazon River dolphin), Lipotidae (Chinese river dolphin) and Pontoporiidae (La Plata river dolphin) have re-adapted to fresh water (Barnes, 1984). Some species of Delphinidae have also adapted to fresh water (occur in both salt and fresh, probably as separate races); including *Sotalia fluviatilis, Orcaella brevirostris* and *Neophocaena phocaenoides.*

The family Delphinidae is the largest and singularly most diverse group of cetaceans, consisting of some thirty-two species falling within seventeen genera. They consist primarily of the smaller toothed whales, many of which are endemically distributed in the various ocean basins throughout the world. The larger cetaceans tend towards a more cosmopolitan distribution (Gaskin, 1982).

Characters used in systematic studies have generally concentrated on aspects of cranial morphology such as the external nasal passages and facial complex (Mead, 1975), the tympano-periotic bones (Kasuya, 1973) and the air sinus system (Fraser *et al*, 1960). More recently, molecular analyses of protein (Shimura *et al*, 1987), nuclear DNA (Schlötterer *et al*, 1991) and mitochondrial DNA (Southern *et al*, 1988) have contributed towards a more comprehensive phylogenetic understanding of cetaceans.

3.3.1 Higher-level Classification of the Extant Cetacea

Phylum : Mammalia

Order : Cetacea

Sub-order : Mysticeti

Families : Balaenidae (Right whale)
Neobalaenidae (Pygmy Right whale)
Eschrichtiidae (Grey whale)
Balaenopteridae (Rorquals)

Sub-order : Odontoceti

Families : Physeteridae (Sperm whales)
Kogiidae (Pygmy sperm whale)
Monodontidae (White whales)
Ziphiidae (Beaked whales)
Delphinidae (Dolphins)
Phocoenidae (Porpoises)
Platanistidae (River dolphins)
Iniidae (Amazon River dolphin)
Pontoporiidae (Franciscana)

(Perrin, 1989)

3.3.2 Family Delphinidae :

Genera : *Peponocephala*
Feresa
Pseudorca
Orcinus
Globicephala
Steno
Sotalia
Sousa
Lagenorhynchus
Lagenodelphis
Delphinus
Tursiops
Grampus
Stenella
Lissodelphis
Cephalorhynchus
Orcaella

(Taken from Perrin, 1989.)

3.4 TAXA SAMPLED IN THE PRESENT STUDY

The present study sampled taxa from eight of the seventeen genera of Delphinidae, these being :-

Genera	Species
Cephalorhynchus	*heavisidii*
Lagenorhynchus	*obscurus*
Tursiops	*truncatus*
Delphinus	*delphis*
Grampus	*griseus*
Feresa	*attenuata*
Globicephala	*melas*
Stenella	*coeruleoalba*

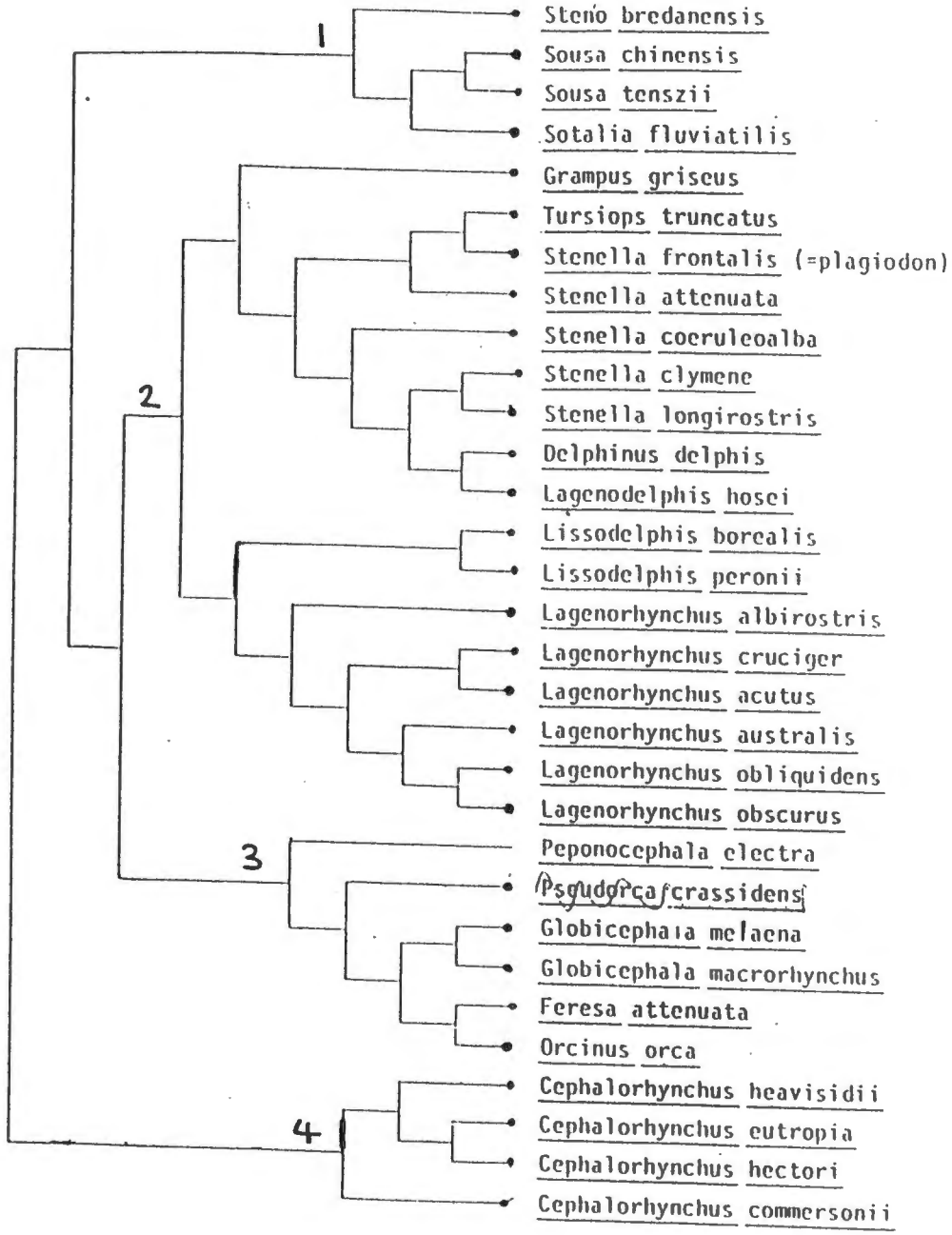


Figure 4 : A tentative phenogram of Delphinidae (based on the literature and personal observations) by Perrin (unpublished).

Note : In a later publication Perrin (1989) puts *Lissodelphis* under sub-family Lissodelphinae.

Key :

1. Steninae
2. Delphininae
3. Globicephalinae
4. Cephalorhynchinae

3.5 MORPHOLOGICALLY BASED GROUPINGS OF PERTINENT GENERA

Ancestral lineages within the family Delphinidae have not been fully resolved, as the variations to these pertinent generic groupings (subfamily status) demonstrate.

3.5.1 Kasuya's Classification (1973)

Kasuya used the tympano-periotic bones as a major morphological differentiating character.

Globicephalinae :

Grampus
Feresa
Globicephala

Sotalinae :

Cephalorhynchus

Delphininae :

Lagenorhynchus
Tursiops
Stenella
Delphinus

3.5.2 Mead's Classification (1975)

Mead's studies are based on the external nasal passages and facial complex.

Orcinae :

Globicephala
Feresa

Cephalorhynchinae :

Cephalorhynchus

Delphininae :

Grampus
Lagenorhynchus
Tursiops
Stenella
Delphinus

3.5.3 Fraser and Purves' Classification (1960)

Fraser and Purves' classification is based on the air sinus system.

Orcinae :

Globicephala
Feresa

Cephalorhynchinae :

Cephalorhynchus

Delphininae :

Grampus
Lagenorhynchus
Tursiops
Stenella
Delphinus

3.5.4 Perrin's Classification (1989)

Globicephalinae :

Feresa

Globicephala

Cephalorhynchinae :

Cephalorhynchus

Delphininae :

Grampus

Lagenorhynchus

Tursiops

Stenella

Delphinus

3.6 MOLECULAR-BASED CLASSIFICATION OF DELPHINIDAE

3.6.1 Allozyme Study

The first extensive molecular systematic study reported on the Delphinidae is an allozyme study by Shimura and Numachi in 1987. Using starch-gel electrophoresis at nineteen genetic loci encoding enzymes, the genetic variability and differentiation of three families (*Beradius* bairdii from family Ziphiidae, eight species of Delphinidae and three species of Phocoenidae) were examined. Of the seven genera sampled, four are the same as those sampled in the present study. The genera sampled are *Peponocephala, Globicephala, Pseudorca, Stenella, Tursiops, Lagenorhynchus* and *Steno*. The results are very similar to morphologically-based classifications.

The Phocoenids are grouped as a separate family. The taxa from *Stenella, Tursiops* and *Lagenorhynchus* are grouped together. *Peponocephala, Globicephala* and *Pseudorca* assume a close genetic relationship.

Difficulty in calibrating allozyme divergence times resulted in estimates of either 3,5 to 5,5 Myr ago or 13,3 to 20 Myr ago for divergence of the three families.

Speciation within the Delphinidae was estimated as beginning at either 2 Myr or 7,6 Myr ago, depending on the statistical technique used to translate genetic variance into time, thus demonstrating the weakness of this method for dating radiation events.

Although rather extended, these dates still correlate reasonably with palaeontologically based points of radiation (see Chapter 2).

(For Figure 5 please see overleaf.)

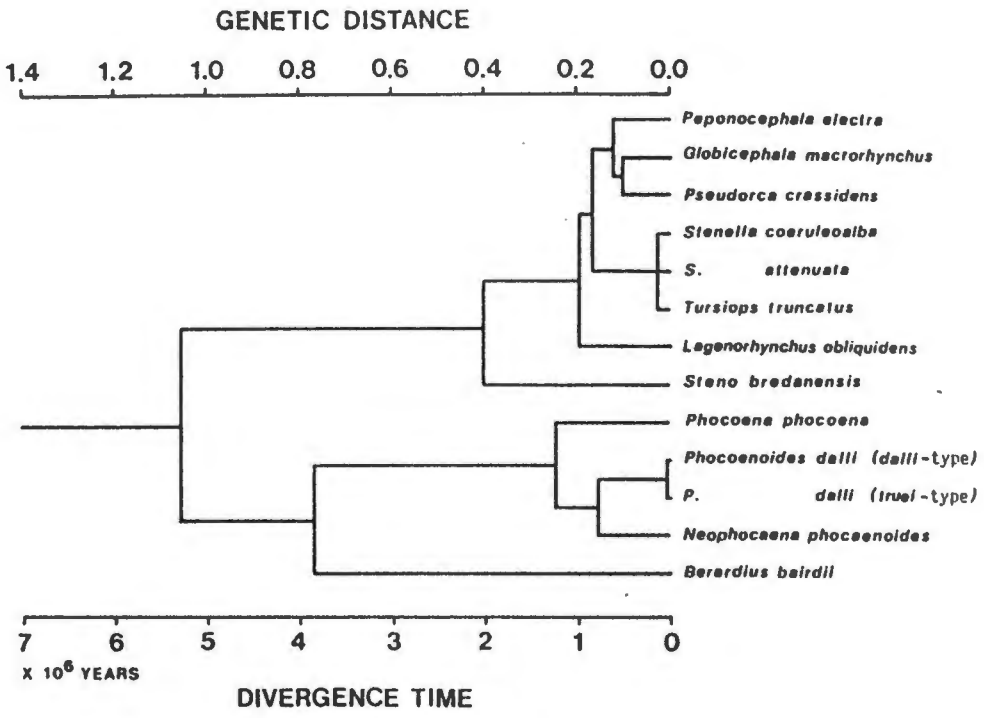


Figure 5 : Biochemical similarity dendogram of toothed whales based on genetic distance. Divergence time by calibration of Nei (1975) is also shown (Shimura *et al*, 1987). Note that the Delphinoidea are not monophyletic.

3.6.2 Nuclear DNA (nDNA) Based Study

In the second molecular study, Schlötterer *et al* (1991) compared a total of 397 nucleotide flanking sequences between eleven species of Cetacea. The four simple sequence loci analysed were located in a non-coding region of eukaryotic (Nuclear) DNA.

Schlötterer *et al's* phylogenetic studies are at a higher level (family and suborder) than the present study's and as such they do not address the generic interrelationships of the four members of Delphinidae sampled.

The most interesting aspect of this study is the very recent dates of radiation obtained between families and suborders when compared with palaeontologically based estimates. The implications and possible explanations of the recent dates will be elaborated upon under "Discussion", as the present study, using mtDNA rather than nDNA and a different molecular technique, obtained very similar palaeontologically discordant results.

CHAPTER 4

MITOCHONDRIAL DNA AND METHODS OF PHYLOGENETIC ANALYSIS

4.1. DESCRIPTION OF MITOCHONDRIAL DNA

The mammalian mitochondrion DNA (mtDNA) is haploid and maternally inherited, an attribute which makes it useful in genealogical studies (Gyllensten *et al*, 1985). It is structurally very stable and consists of a circular double stranded DNA molecule of about 15000 - 17000 base pairs (bp). The mtDNA's small size makes it very suitable for restriction enzyme based analysis or DNA sequencing techniques. The mitochondrial genome is about 1/10000 the size of the smallest animal nuclear genome (nDNA) (Li *et al*, 1990), and also consists of multiple copies per cell.

Advantages of mtDNA over nDNA are -

(a) mtDNA is easily extracted and purified from post-mortem tissue;

(b) its small size renders it amenable to restriction enzyme analysis;

(c) it accumulates point mutations at a rapid rate; and

(d) it is haploid and does not undergo recombination.

(Hewitt *et al* (eds), 1990).

MtDNA can be relatively easily isolated and purified as mitochondria -

(a) yield large amounts of DNA;

(b) have a high copy number; and

(c) occur in an organelle rather than in the nucleus.

The mtDNA consists of non-repetitive sequences of thirteen protein coding genes, two rRNA genes, twenty-two tRNA genes and a control region that contains sites for replication and transcription initiation (Anderson *et al*, 1981).

The mitochondria are commonly known as the "power plants" of the cell, as it is there that the oxidation of carbohydrates is completed and where most of the adenosine triphosphate (ATP) is produced (Darnell *et al*, 1986).

4.1.1. Base Substitution Rate of mtDNA

Even though the mitochondria perform the vital function of oxidation of carbohydrates in the cell, the evolution of their DNA is five to ten times more rapid than that of single copy nuclear DNA (scnDNA) (Brown *et al*, 1979). These authors calculated the mean rate of sequence divergence of the mtDNA genome to be 2% per million years (or 1% base substitution rate) for mammals. This estimate was calibrated using fossil records and protein based datings.

The mtDNA's high base substitution rate makes it more sensitive than nDNA for lower level (species or genera) phylogenetic studies.

Single base changes, deletions, insertions or inversions of one or more bases, as well as nucleotide rearrangements are types of mutations which contribute towards evolution of DNA. Single base changes predominate in mtDNA. Transition (pyrimidine to pyrimidine or purine to purine) occurs far more frequently than transversion point mutations (pyrimidine to purine or vice versa) (Watson *et al*, 1987).

(For Figure 6 please see overleaf.)

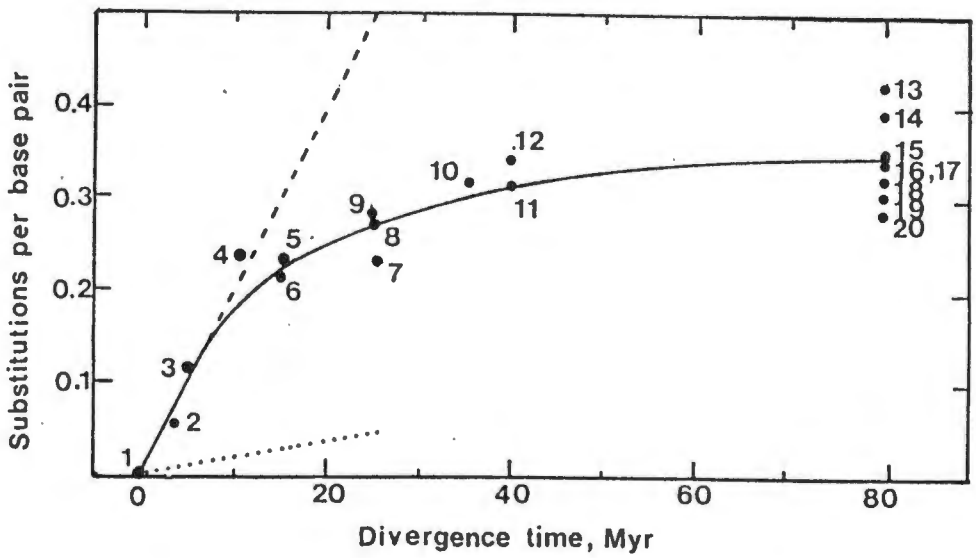


Figure 6 : Dependence of sequence divergence in mtDNA upon time of divergence (Taken from Brown *et al*, 1979)

The y-axis shows the estimated number of base substitutions that have accumulated per base pair (p) for each species compared. This number is calculated from restriction map comparisons by use of equations 1 and 3 in Brown *et al* (1979). The rate of substitution for mtDNA is obtained from the initial slope of the curve, indicated by the broken line. The rate for single copy nuclear DNA (scnDNA) is obtained from the slope of the dotted line. Each point on the graph corresponds to a comparison of 2 species and of individuals within species.

1: mean difference among humans
2: goat and sheep
3: human and sheep
4: baboon and rhesus
5: guenon and baboon
6: guenon and rhesus
7: human and guenon
8: human and rhesus
9: human and baboon
10: rat and mouse
11: hamster and mouse
12: hamster and rat
13-20: rodent-primate species pairs

Both fossil and protein data were used to estimate the times of divergence.

4.2. METHODS OF MITOCHONDRIAL DNA SEQUENCE ANALYSIS

The methods to be briefly discussed are :-

(a) restriction fragment length ploymorphism (RFLP);

(b) endonuclease restriction site mapping (RSM); and

(c) sequencing.

4.2.1 Restriction Fragment Length Polymorphism (RFLP)

This is the quickest and simplest method of mtDNA characterisation, but it can only be used for groups of closely related taxa (for example at the sub-species level or for within-population studies) as once the proportion of shared fragments becomes low this method becomes inaccurate (Harley, 1988), (Hillis *et al*, 1990).

The basis of this method is as follows :-

The restriction enzyme cleaved mtDNA fragments from the chosen taxa are separated electrophoretically and their molecular weights (mw) are calculated. It is assumed that comparative taxa sharing fragments of the same mw will also share flanking cleavage sites. An enzyme's cleavage site(s) are characterised by a specific nucleotide sequence, usually of four or six base pairs in length. Using a formula developed by Nei *et al* (1979) the sequence divergence (between the selected taxa), based on the proportion of shared fragments can be calculated.

To calculate the position of shared sites (S) :

$$S = \frac{2 \times \text{No of shared fragments}}{x\ ty}$$

where x = No of sites in Taxon 1
y = No of sites in Taxon 2

To calculate sequence divergence (d) :

$$d = -\log \frac{S}{r}$$

where r = No of nucleotides in the restriction cutting sequence

The possibility of convergence, the occurrence of which is more probable as the degree of relatedness between comparative taxa decreases, constitutes a source of error. As already stated, it is assumed that similar mw fragments between taxa is indicative of shared flanking cleavage sites. Convergence is the possibility that fragments of similar mw are produced by different cleavage sites (Hillis *et al*, 1990).

Restriction fragment length polymorphism or variations thereof have been used extensively in cetacean population studies, as described in the special issue (No 13) Report of the International Whaling commission (Hoelzel (ed), 1991).

4.2.2 Restriction Site Mapping (RSM)

Restriction site mapping entails the plotting of the selected restriction endonucleases' cleavage sites on the genome. This is essentially achieved through analysis of single and double digest fragment size data, a process which is described in detail in subsequent chapters (Nei *et al*, 1979).

Usually 45 to 50 restriction sites are plotted on the mtDNA genome. Comparative analysis of restriction site maps based on the proportion of aligned sites enables the sequence divergence to be calculated. In addition the sites themselves can be used as characters in cladistic analysis.

Providing the enzymes' cleavage sites are accurately plotted on the mtDNA genome and provided that a large enough sample is used (approximately 50 sites), the restriction site mapping technique offers high resolution for phylogenetic analysis at the species and genus levels.

4.2.3 Sequencing

With the advent of the Polymerase Chain Reaction (PCR) technique, sequencing is now probably the most commonly used method of DNA sampling. The PCR technique enables the targeted segment of nDNA or mtDNA to be amplified or multiplied sufficiently for direct sequencing, a process which used to require much more laborious cloning techniques.

Ideally, single stranded DNA should be used for dideoxy sequencing. This is achieved by amplifying the targeted sequence initially with two and subsequently with a single oligonucleotide primer. A number of PCR cycles then results in the generation of an excess of one strand of DNA (Hewitt *et al* (eds), 1990).

Sequencing, which entails the reading of the order of nucleotides on a selected region of the DNA genome, provides direct and objective data for distance method based or cladistic analyses (Harley, 1988).

In comparative studies the same segment of nDNA or mtDNA should be used as the base substitution rate varies over different genomic regions. Non-coding regions (such as the D-loop region in mtDNA) evolve faster than coding regions (Li *et al*, 1990), and hence are a more sensitive measure of sequence divergence, and are most appropriate for studies on closely related taxa.

Neutral or non-coding genomic regions evolve relatively independently of selective pressure (Harley, 1988).

A possible source of error is that only a fragment of the total genome is actually used. The usually relatively small sample (number of bases to be sequenced) compared to the total genome length, as well as the question of whether the selected region(s) to be sampled are representative of the sequence divergence of the complete DNA genome, are error factors which must be considered when using this method for estimating the degree of sequence divergence between taxa.

CHAPTER 5

THEORY OF THE RESTRICTION ENDONUCLEASE SITE MAPPING TECHNIQUE

5.1. INTRODUCTION

A restriction endonuclease (REs) recognises specific sequences of nucleotide pairs, usually between four and eight bp in length, and cleaves the DNA at that position. Digestion of the mtDNA (which has a circular conformation) with a REs will result in the cleavage of the mitochondrial genome into the same number of fragments as there are restriction sites. The more similar the mtDNA sequences, the more similar the DNA "characterising" cleavage positions will be, as the number and location of restriction sites vary with nucleotide sequence (Nei *et al*, 1979). The proportion of shared restriction sites in between taxa comparison is expected to decline as their DNA sequences diverge. The construction of restriction endonuclease site maps entails the plotting of a number of enzymes' restriction sites on the mtDNA genome. The resultant restriction site map is a method of sampling the whole mitochondrial genome, or it can also be conceived of as a "characterisation" of the mtDNA. Inferred phylogenies are constructed from a comparative analysis of a number of such restriction site maps (using either cladistic or distance methods).

5.2 DOUBLE DIGEST FRAGMENT ANALYSIS

Restriction site mapping is based on the fragment analysis of single and double digests. Single digest data provide the number of restriction sites, but not their positions, on the mtDNA genome. The positions of a restriction enzyme's (REs) sites can be deduced from a computational fragment analysis of single and double digest data obtained from at least three different enzymes (three-way analysis). Re-phrased, a REs site can only be calculated with reference to at least two other enzymes. This is achieved by undertaking all three possible double digest combinations of the three selected enzymes, and the subsequent reconstruction of the complete genome from the best possible fit of the cleaved fragments.

To avoid computational error it is preferable to keep double digest combinations as simple as possible. Using one to four cutting enzymes in double digests will, depending on the combinations used, produce a manageable double digest of between two and eight fragments. An enzyme that cuts many times is ideally mapped in using a series of single or double cutting enzymes whose positions have already been mapped and which are comprehensively spread over the whole genome.

In general, an enzyme's sites can only be plotted with reference to the (immediate) flanking sites of two other enzymes. There are a number of sources of error variance which can confound accurate computations (refer Chapter 6). A calculated REs cleavage site is a best estimate of its actual position.

5.3 RESTRICTION ENZYME SITE MAPPING PROCEDURE

5.3.1. Decide on single and double digest enzyme combinations to be used.

5.3.2. After using the enzymes to cleave the mtDNA, sort the fragments using gel electrophoresis and visualise using the end-labelling technique.

5.3.3. Measure the molecular weight (mw) of the fragments. The mw of the fragments are measured using a calibration curve based on a sample with fragments of a known size run on each gel (Hillis *et al*, 1990). Such a calibration curve is constructed on the basis of the direct function between the mw (measured in base pairs) and the distance migrated (measured in millimetres) of the mtDNA fragments (see Fig. 9). The present study used Lambda DNA ( ) digested with Hind III to produce eight fragments of known mw (refer Appendix III for number and mw of marker fragments).

5.3.4. Calculate the REs sites on the mtDNA genome using the fragment size data thus obtained in the three-way analysis system.

5.4 RESTRICTION SITE MAP CONSTRUCTION USING THE 3-WAY ANALYSIS METHOD

Example : Three-way analysis of REs Sac II, Asp 718 and Eco RV, using an ideal data base set

Given : Total mt. DNA genome length (16400 base pairs)

Single digest fragment sizes :-

Sac II	1.	14 700
	2.	1 700
Asp 718	1.	11 950
	2.	4 450
Eco RV	1.	16 400

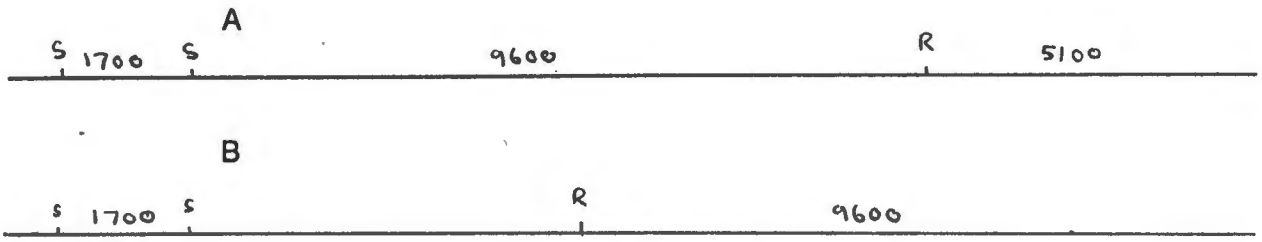
Double digest fragment sizes :-

Sac II & Asp 718	1.	11 100
	2.	3 600
	3.	850
	4.	850
Sac II & Eco RV	1.	9 600
	2.	5 100
	3.	1 700
Asp & Eco RV	1.	5 975
	2.	5 975
	3.	4 450

Assume that maps are orientated, i.e. ignore possible inverted solutions. In actual construction, REs maps are aligned on two Sac II sites which are invariant through almost all the Vertebrata and are orientated using a similarly invariant Hpa I site. (Sac II positions are 676 bp and 2356 bp; Hpa I position is 5540 bp, that is, 3184 bp to the right of the second Sac II site.)

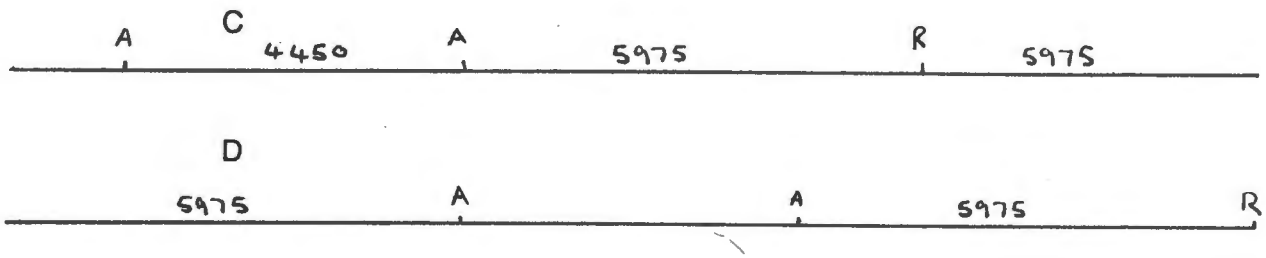
Sac II and Eco RV

There are two possible reconstructions which satisfy the single and double digest data :



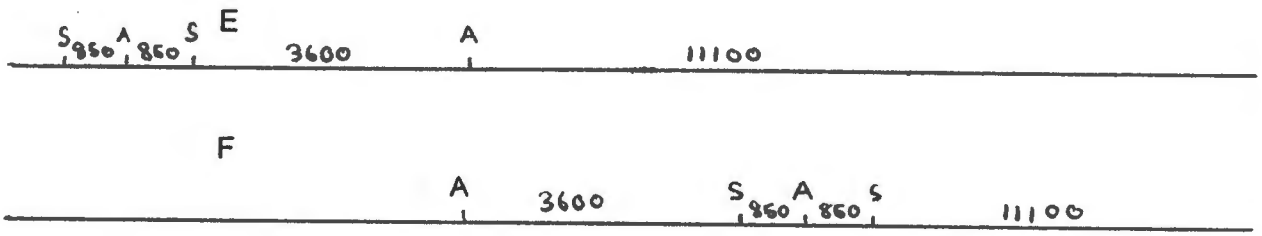
Asp 718 and Eco RV

Possible reconstructions :



Sac II and Asp 718

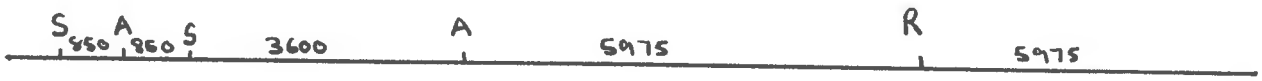
Possible reconstructions :



Eco RV's position is calculated using Sac II and Asp 718. Of the possible solutions, maps A and C's calculated position of Eco RV correlate the highest (the example is using an ideal data base set, therefore the correlation will be 1):

Asp 718's restriction sites are calculated using enzymes Eco RV and Sac II. Maps C and E correlate Asp 718's position.

The final map which reflects the highest inter-correlation between all three enzymes' calculated positions is a composite of maps A, C and E, which is :-



The three-way analysis method computes all possible fragment fits of the three double digests and then constructs a final map on the basis of the highest degree of correlation between the three enzymes' calculated positions, with reference to two other enzymes.

For the initial map construction it is preferable to use simple cutting enzymes which produce unambiguous double digest data. The calculated positions of these enzymes should be as accurate as possible as subsequent enzymes will be mapped in with reference to them.

5.5 3-WAY ANALYSIS: A SOLUTION TO UNFIXED SITES IN DOUBLE DIGESTS

An enzyme's sites can only be plotted with reference to the (immediate) flanking sites of another enzyme. Unfixed sites in double digests are those sites which are not flanked by the other enzymes' cleavage positions. Therefore such sites cannot be plotted using one set of double digest data. The 3-way analysis solves the unfixed site problem by introducing a third enzyme which restricts the unfixed fragment, thereby enabling its position to be calculated with reference to the two other enzymes.

For example :

A'$_1$'s position can be either -

or

The inclusion of a third enzyme (in a 3-way analysis) will fix A'$_1$'s site:

5.5.1 3-Way Analysis: solving the problem of multiple solutions for simple two enzyme double digests

Analysis of certain single and double digest fragment size data can lead to a number of possible solutions. As demonstrated under Section 5.3 (restriction enzyme site mapping procedure), only one such solution will work in the 3-way analysis, or at least will resolve at a lower error level.

5.6 CONCLUSION

Restriction site mapping is time-consuming, even with the assistance of the Resolve computer program (refer to Appendix III). In the present study, 100 to 150 single and double digest reactions were needed to map the sixteen chosen enzymes per sample. The sixteen enzymes mapped had on average forty-five to fifty combined recognition sites or, translated into base pairs (6 bp's per site; total genome length estimated at 16400 bp's), about 1,8% of the total mtDNA genome was sampled.

Contrary to the sampling method of sequencing (which entails "reading" the order of nucleotides on a specifically selected region of the genome), restriction site mapping draws samples (in the form of recognition sequences) from the whole (mitochondrial) genome. Restriction site mapping therefore samples both coding and non-coding genomic regions of the mtDNA.

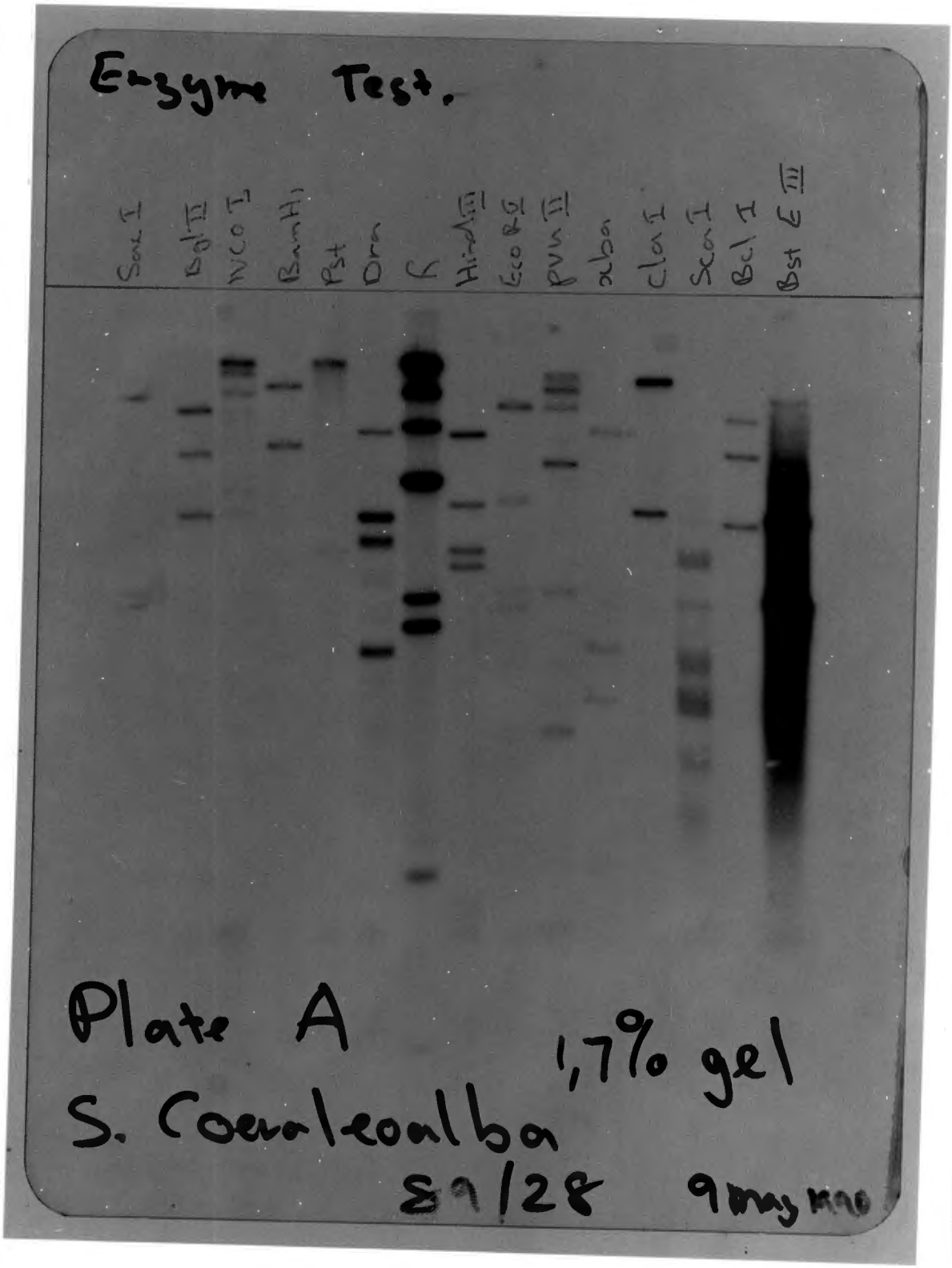


Figure 7 : An autoradiograph showing the number of cleavage sites (as determined by the number of fragments) of fourteen restriction endonucleases (REs) on the mtDNA genome of *Stenella coeruleoalba*

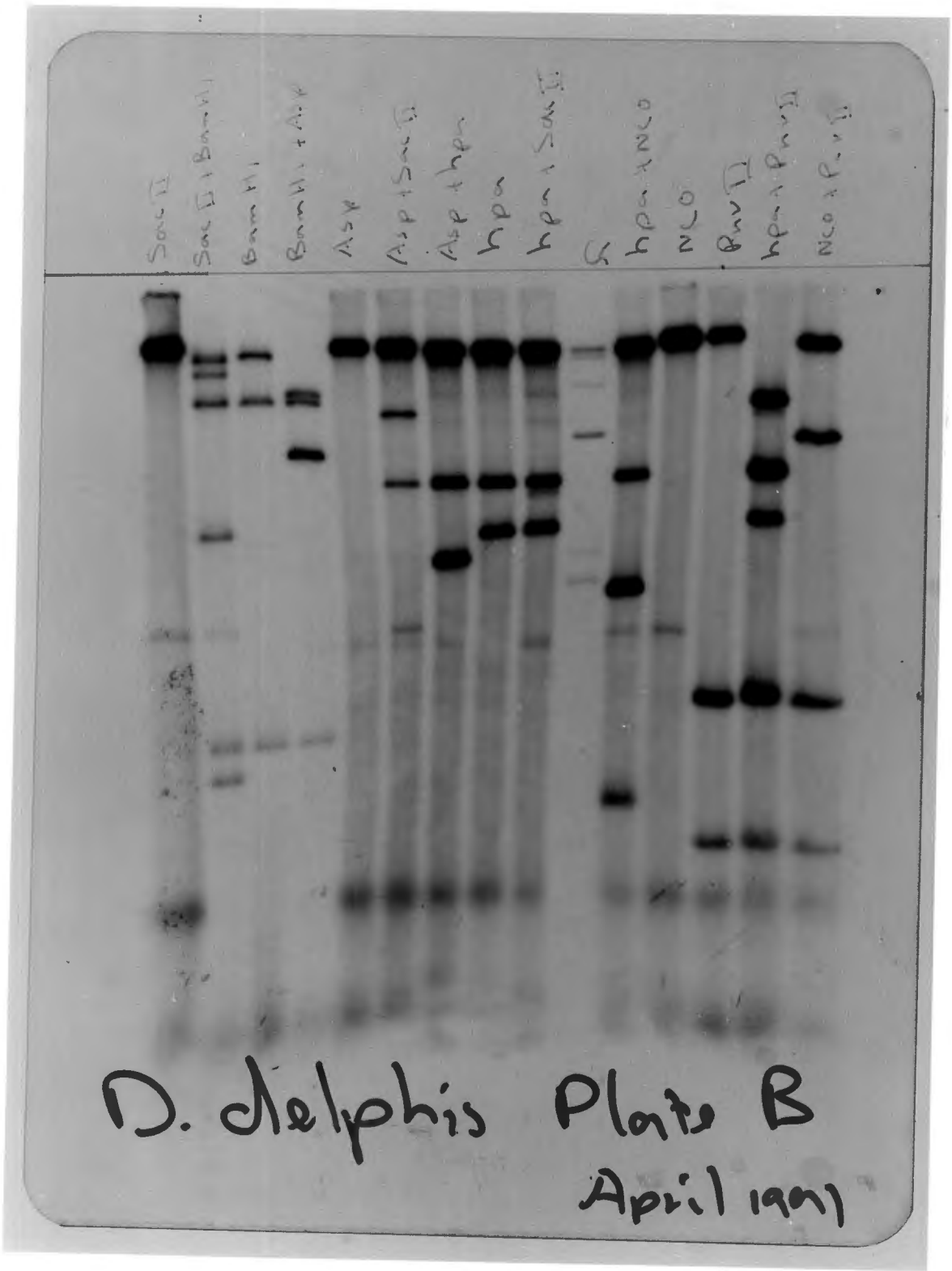


Figure 8 : An autoradiograph showing fourteen single and double digest combinations (*Delphinus delphis* mtDNA).
The number and size (molecular weight) of REs cleaved mtDNA fragments from such combinations are used in the 3-way analysis method to determine the REs cleavage site(s) on the mtDNA genome. The molecular weights of the cleaved fragments are calculated using the known fragment sizes of phage Lambda (λ) DNA restricted with Hind III.

CHAPTER 6

CRITICAL ANALYSIS OF THE CONSTRUCTION OF RESTRICTION SITE MAPS OF MITOCHONDRIAL DNA

6.1. INTRODUCTION

The ideal molecular data base set for comparative analysis between taxa would be the complete sequences of the genomes themselves. The extremely time-consuming and laborious nature of such a task (as mentioned in the Overview) makes it necessary to sample the genome instead. The construction of restriction site maps is just such a sampling technique. Theoretically the basis of this method is quite simple. It entails the plotting of the restriction sites of the chosen enzymes on the mitochondrion's genome. However, these positions have to be deduced primarily from the single and double digest fragment-size data obtained from the autoradiographs and from an, at times, complex computational methodology. An inherent weakness of restriction map data is the indirect (as opposed to sequencing's direct) representation of changes on the mtDNA genome. The term "indirect" implies a degree of error in calculation of sequence data from restriction site changes, as well as error variance in the methods used to calculate the restriction sites' positions.

The following section offers a critical discussion of the computational aspect of the restriction site mapping technique. It also describes possible sources of error variance and how they can be contained.

6.2 ACCURACY AS A FUNCTION OF SITE ALIGNMENTS

Harley *et al* (m.s. in preparation) have noted that "... overall accuracy (of plotted restriction sites) is a function of the accuracy of site alignments" and again that " ... a subjective assessment of aligned sites on comparative maps could be a major cause of variation". The cornerstone of this sampling method is the accurate calculation and plotting of the restriction sites on the genome. However, certain sources of error variance can confound such computations. Restriction maps whose site positions are subject to a high degree of error are porportionately less valid, and are thus of less use for comparative analysis. It is probably a fair statement that any restriction site map, no matter how carefully constructed, contains at least some degree of error variance.

Harley *et al* have attempted to curb error variance, most notably in the form of "subjective assessment of aligned sites" by constructing the maps independently of each other. In any subsequent comparative analysis between maps, restriction sites falling, usually, within 1% of total genome length are taken as being similar. "Subjective assessment of aligned sites" can be defined as being a personal evaluation, or the bias of the analyst, in the concurrent construction of restriciton site maps, to include or exclude sites when assessing site alignments, or the "shifting" of restriction sites so that they align more "reasonably".

6.3 RESOLVE: A COMPUTER PROGRAM DESIGNED TO FACILITATE RESTRICTION SITE MAPPING

To further reduce human sources of error variance, Harley (m.s. in preparation) wrote a computer program (Resolve Version 2.7) which, amongst other numerous functions, calculates, using single-digest and double-digest fragment size data for at least three enzymes, the best probable "fit" of the double-digest fragments, thereby computing the most likely positions of the restriction endonuclease sites within a pre-selected error variance, or that giving the least error variance.

Although the program greatly facilitates the computations involved there are still problematical areas, one of which is the error in size estimation. There is a direct relationship between percentage error and molecular weight of the fragments. Therefore the actual error variance (i.e. molecular weight or number of base pairs) will be much greater for large fragments than for small fragments at the same percentage error. For example, a large fragment of 8000 base pairs at a 2% error variance will have a molecular weight error variance of 160 base pairs, whereas a fragment of 1000 molecular weight will only have a molecular weight error of 20 base pairs. This can prove problematical in double-digest computations, which consist of both large and small fragments, as the computational base pair error variance for the small fragments will be much smaller than for the larger fragments.

The relative error variance between large and smaller fragments is quite independent of error variance inherent in a site's calculated position (relative to its actual position) on the genome, where calculated percentage error is a function of the genome's complete size. Using the same example, the computer may reject a possible "fit" of fragments because of the extremely small base pair error variance of the small fragments, whereas in real terms (that is with reference to the actual positions of the site on the genome) the rejected site position calculated from this might be extremely close.

A final limitation of the program is its inability to compile solutions from double-digest data whose composite fragments do not "lock in" or are not "fixed" in a direct relationship with each other. (An enzyme's site positions are calculated using the immediate flanking positions of a previously mapped enzyme [refer Chapter 5]). This is not a fault of the program, but rather a limitation of the raw data itself. Multi-cutting single restriction endonuclease site positions have to be gradually mapped in by using a number of simple cutting enzymes in a series of double-digest reactions.

The above discussion implies that it is not feasible to blindly use the computer program for double-digest calculations. It is essential to have a working knowledge of both the computational aspects of mapping and at least some knowledge of the arithmetic logic behind the program itself. The program can greatly facilitate calculations if used as an accessory to one's working knowledge of the technique.

However, calculation of site positions is only one function of the Resolve program, the whole of which is quite indispensable when used for data storage and management, visualisation of site alignments, comparative analysis of restriction site maps, in transforming molecular data into statistically acceptable characters and as an interface between molecular data base sets and comparative analysis programs such as the various distance and cladistic measures. (See Appendix III for working details of Resolve 2.7.)

6.4 SOURCES OF ERROR IN THE CONSTRUCTION OF RESTRICTION SITE MAPS

6.4.1 Translation of Distance Migrated into Molecular Weight

The measurement error which occurs during the translation of distance migrated into the fragment molecular weight (mw) is a function of both gel density and fragment size.

6.4.1.1 Fragment size

Measurement of large fragments is inaccurate. The larger the fragment size the more inaccurate the translation of distance migrated into molecular weight. This is due to the scale of the graph in which the mw increases logarithmically in relation to a linear distance migrated scale (see Figure 9).

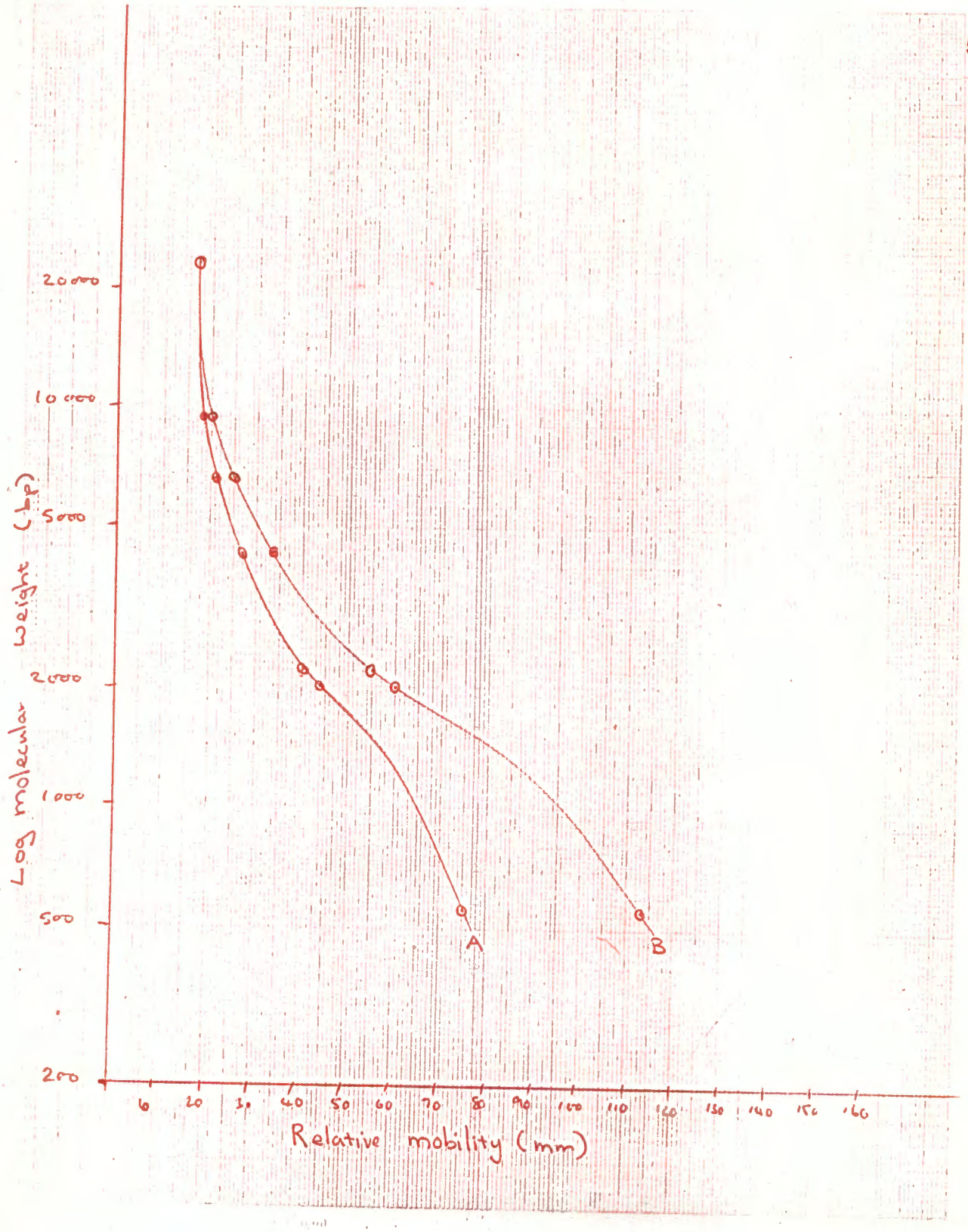


Figure 9 : Semi-logarithmic plot of the relative mobilities (mm) and molecular weights (bp) of restrictive fragments resulting from a Hind III digestion of phage Lamda DNA ( ).

Symbols used :

A 2,0% agarose gel

B 1,5% agarose gel

6.4.1.2 Gel density

Higher density agarose gels will have the effect of increasing the ratio between mw and distance migrated, since the mtDNA fragments do not migrate as far in higher density gels as they do in lower density ones. This increased ratio is reflected in the steepening of the line function of the graph and hence the more inaccurate translation of distance migrated into mw (see Figure 9). However, if the fragments are electrophoresed for longer the resolution will improve. In general, higher density agarose gels (2,0%) are used to measure the mw of small fragments (< 1000 bp) and lower density gels are used for a more accurate measurement of large fragments (> 4000 bp).

6.4.1.3 Quality of agarose gels

Figure 10 shows the retarded migratory rate of Lambda in the extreme left-hand lane when compared to the second Lambda on the right. (The broken bands of the left-hand Lambda indicate an impeded migratory pattern.) Altered mtDNA fragment migrations can affect accurate translation of distance migrated (mm) into molecular weight (bp), especially for the larger fragments. For example, the difference in distance migrated of the fourth Lambda band is 1,75 mm or 300 bp between the two Lambdas.

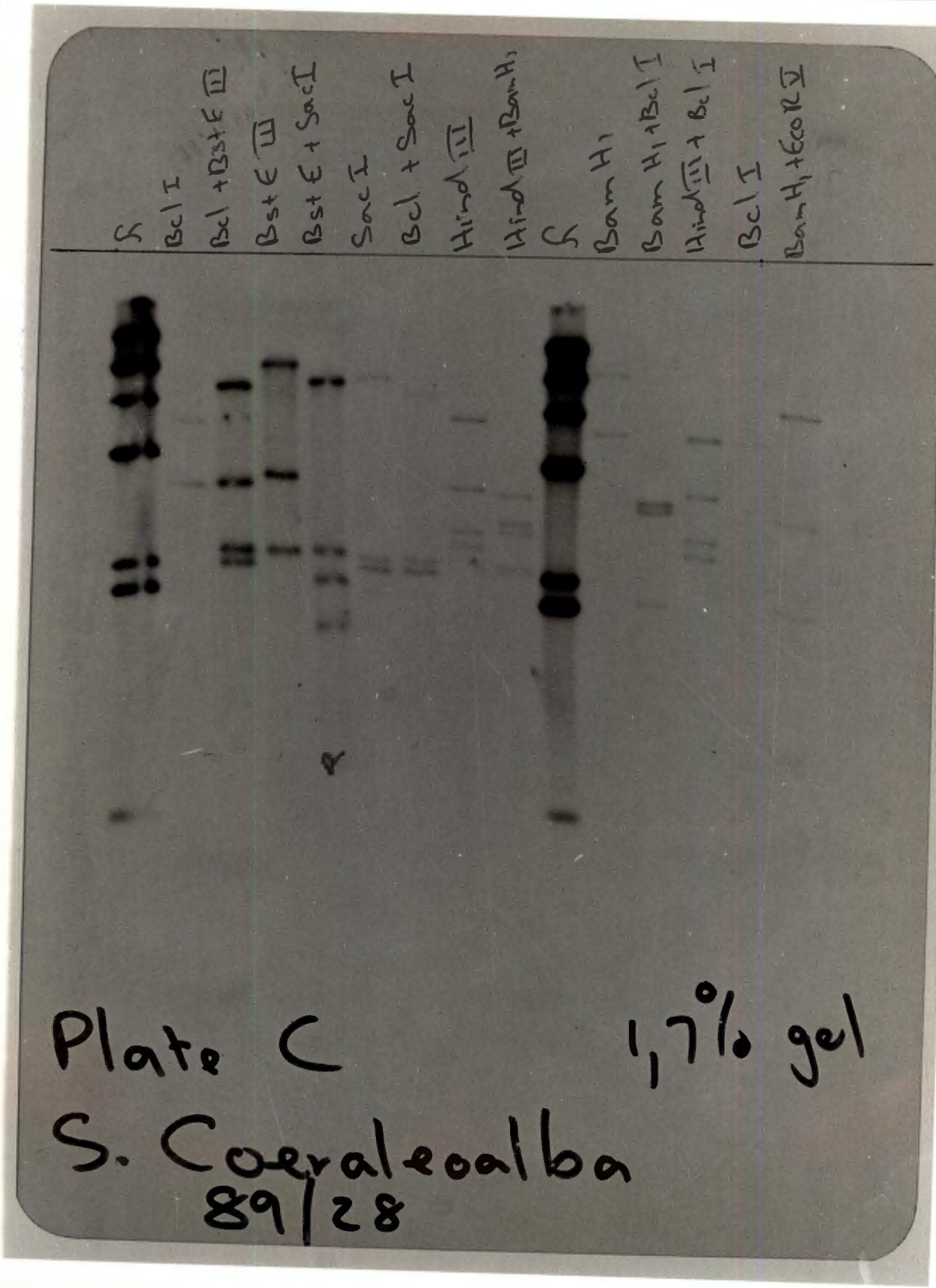


Figure 10 : An autoradiograph showing the different rates of migration of phage Lambda cleaved with Hind III, due to a poorly made agarose gel which has impeded the migration of the left-hand Lambda in comparison with the Lambda on the right side of the autoradiograph. The left-hand Lambda's broken bands are an indication of the impeded migrations of the phage's fragments. Such altered migratory rates can affect the correct translation of (fragments) distance migrated (mm) into molecular weight (bp), especially for larger fragments (refer Section 6.4.1.3).

6.4.2 Indirect Measurement of Large Fragments

The mw of the larger fragments can be deduced by either subtracting the mw of the smaller fragments from the total genome length or by cleaving them in a double digest reaction. Subtracting the mw of the small fragments from the total genome length is also fallible as the measured smaller fragments will have at least some degree of error, and there is always the possibility (in single digests) that there are "lost" fragments (that is, fragments that remain undetected because of their very low mw or from a poor quality autoradiograph), whose mw is invariably added to the large fragment so that the total mw of the fragments equals the genome's size. A more feasible way to calculate a large fragment's mw is to cut it in a double digest reaction. The summation of the composite fragments' mw will provide a more accurate estimate.

6.4.3 "Lost" or "Missing" Fragments

Theoretically, fragments of at least 100 bp can be separated on an agarose gel and visualised on the autoradiograph. Practically this is not always achieved, as such fragments may migrate off the gel, or they remain undetected due to band spreading or nuclear DNA contamination, which is frequent in the lower mw region.

Complex double digest patterns of eight to ten bands are common. On occasion fragments which are known to exist (as the total number of fragments of two single enzymes must equal the number of double digest fragments) are not visualised on the autoradiograph and are designated as "missing". They are usually small fragments, and their estimated size must be deduced by analysing the known double digest fragments, an endeavour which can lead to misconstrued deductions. Overlapping or double bands are the result of fragments of similar mw being produced by different cleavage sites. The presence of a probable double band is deduced by a missing fragment of similar mw to an existing one. A double band is also usually darker and broader than a single band.

6.4.4 Partial Digestion

Partial digests are the result of incomplete digestion in which not all the mtDNA molecules are cleaved at all sites (Hillis *et al*, 1990), resulting in the presence of extra bands. Partial digestions are detected by the total size of fragments exceeding the genome size beyond acceptable error. It can prove problematical to locate the "partial cuts", especially if they are present in double digest banding patterns. In such a case it is better to repeat the digest using either an increase in the amount of enzyme used or by using a new batch of restriction enzymes.

6.4.5 Satellite DNA

Highly repetitive nuclear DNA sequences are known as satellite DNA (Freifelder, 1983). Satellite DNA was visualised on many of the autoradiographs (of different species) and occurred in the 1800 bp region. Satellite DNA can be recognised by its appearance in the same position with different enzymes (i.e. NCO I, Hind III and Asp 718). The presence of satellite DNA can render double digest fragment analysis more difficult, especially when using RE Sac II, which has a 1820 bp fragment which is in a similar position to the satellite band. (The satellite DNA band was not evident in single Sac II digests.) (Refer Figure 8, which shows the satellite band just below the Sac II 1820 band.)

6.4.6 Cumulative Error

A major criticism of the restriction site mapping technique is the cumulative error associated with calculated site positions. Invariably there will be some error accrued in the plotting of the enzymes' restriction sites on the genome. The sites of subsequently added enzymes are calculated using the initial enzyme's cutting positions as a reference (as described earlier in the three-way analysis method). Thus any substantial error accrued in the calculation of the initial enzyme's cleavage sites can affect subsequent computations. This could lead to an inaccurate map, which will make it more difficult to map new enzymes.

The importance of accurate mapping cannot be over-emphasised. For example, even if an enzyme's sites are plotted to within 2% of the actual positions in mtDNA, their positions could differ from the correct ones by as much as 300 base pairs !

As mentioned earlier, in any comparative analysis between maps such as is undertaken for either cladistic or distance based phylogenetic analysis a percentage error is chosen, the degree of which determines whether or not sites are aligned and thus designated "similar". There is no fixed or accepted rule which determines the degree of error to be used when assessing degree of site alignments in between map comparisons, and no discussion on this topic has appeared in the scientific literature. In the final dendograms different topologies can be produced if varying percentage errors are used. Selecting the error variance for between map comparisons should be a function of the validity of the individually constructed maps. The problem is how to measure the validity of calculated site positions if the actual positions are unknown !

As will be described, there are indirect means to validate an enzyme's calculated restriction site positions. On the basis of these checks a reasonable estimation of the degree of error variance can be made and is recommended to be used as the degree of error for between map comparisons.

6.4.7 Gain and Loss of Restriction Sites

There are four types of mutation that can change a RE's cleavage site. These are sequence rearrangements, base substitutions and the addition or deletion of nucleotides within a recognition sequence.

Statistically it is more probable for a cleavage site to be lost than for it to be regained, as to lose a site there need only be one change in the recognition sequence. To regain the site, i.e. for the changed base to revert back to its original type, is only a one in four probability (there are four possible base substitutions) (Hillis *et al*, 1990).

Secondly, base changes do not occur with equal probability as there is a high transition bias (which is greater than 90% in mtDNA). A transition is that type of base change which occurs when a purine or pyrimidine changes into its own type (i.e., from adenine to guanine or thymine to cytosine). A transversion occurs when a purine changes to a pyrimidine, or a pyrimidine into a purine (Watson *et al*, 1987).

6.5 MEANS OF CONSTRUCTING MORE ACCURATE RESTRICTION SITE MAPS

An awareness of the possible sources of error variance when analysing fragment size data (used in conjunction with the program Resolve's excellent data management and computing facilities) is an important requirement for producing accurate restriction maps.

6.5.1 General Rules

These include maximising the purity of the mtDNA, the use of the correct end-labelling technique, improving the quality of autoradiographs and, in the case of ambiguous banding patterns, the use of a series of different combination double-digests. Sometimes it is useful to use two different gel densities for the fragment separation of the same double-digest reaction. Using a higher density gel (2% to 2,2%) will assist in identification and sizing of the smaller fragments, whereas a low density one (1,2 % to 1,5%) permits a more accurate measurement of the larger fragments.

6.5.2 3-way Analysis System

The 3-way analysis system (whether done manually or using the computer) by the very nature of its construction reduces the incorrect computation of restriction sites.

Essentially an enzyme's sites are calculated with reference to site positions in two other enzymes, with each of the double-digest's solutions verifying the calculated positions of the other. Manual computation of a 3-way double-digest series compared with the computer's similar calculations is also a useful way of verifying deduced site positions.

6.5.3 Independent Construction of Maps

Harley *et al*, as mentioned earlier, suggest the independent construction of maps to avoid the "subjective assessment of aligned sites". This method does not make the individual maps more accurate, but it does create a more objective data set for subsequent phylogenetic analysis.

6.5.4 Concurrent Construction of Maps

Under certain circumstances it can be justifiable to construct the individual maps concurrently. This can be done if taxa have an identical fragment pattern with a restriction enzyme and provided the enzyme cuts more than once. (A single fragment could be given by the enzyme cutting at any position around the circular genome.) This can facilitate accurate computation of an enzyme's restriction site positions, as well as being a means of substantiating their calculated positions.

A good approach is as follows :- Work with three to five taxa simultaneously. Run enzyme test gels for all taxa and select initially those enzymes that share identical restriction sites between taxa. This can be easily achieved by running a single enzyme digest of the same enzyme for the different taxa, on the same gel. Identical banding patterns on the autoradiograph are indicative that the mtDNA genomes samples have been severed into fragments of the same molecular weight, from which we may conclude that the tested enzyme has identical restriction sites on the mtDNA genomes of the taxa under study. Therefore the calculated position of the enzyme's restriction sites should also be identical between the taxa.

If an enzyme gives an identical fragment pattern in two (or more) taxa, then a concurrent approach may be appropriate. As mentioned, an identical fragment pattern given by an enzyme for different taxa is indicative of identical restriction sites among those taxa. A means of validating an enzyme's calculated sites then, would be to compare the calculated site positions between taxa that share an identical fragment pattern. If correct, the calculated sites should be the same among the taxa, as it has been previously established (from their identical fragment patterns), that the taxa have identical sites for that enzyme. Enzymes for 3-way double digest reactions can be so selected that, when used for the construction of restriction site maps, they can also validate the calculated sites.

An example will help illustrate this :

	TAXA		
	A	**B**	**C**
Double digest combinations :	Hpa & SacII Hpa & Asp Sac II & Asp	Hpa & Eco R V Hpa & Asp Eco R V & Asp	Hpa & Bcl Hpa & Sac II Bcl & Sac II

From the above it is evident that Hpa's computed positions can be validated from the different double-digest computations of A, B & C taxa. Sac II's positions can be similarly substantiated by comparing the double-digests of A and C. Any gross computational error will be reflected in a low correlation between the calculated site positions of identically cutting enzymes.

Secondly, it is probable that there will be minor differences in the calculated site positions between the taxa. The mean of the calculated positions can be taken as the best estimate of the restriction site's true position.

Such a method of data fragment analysis should ensure accurate computation of the "foundation" enzymes, which are enzymes that form the basic construct of a restriction site map. Such enzymes generally include those that cut one to three times in the mtDNA in question.

Once this has been achieved, enzymes giving more complex patterns and large numbers of fragments can be mapped in, using two foundation enzymes as the reference enzymes in double digestion and subsequent 3-way analysis.

When constructing maps independently, calculated restriction sites can be validated by using different reference enzymes in two (independent) 3-way analyses. If correctly computed, the enzyme's calculated positions should be the same for both approaches. Similarly, ambiguous double digest data (i.e. unresolved partial solutions), which preclude a feasible 3-way analysis solution, should be re-examined in the light of further double digest reactions based on different reference enzymes.

6.6 CONCLUSION

Even if the data from double digest fragment patterns give multiple two-enzyme mapping solutions (and they usually do), the subsequent 3-way analysis frequently gives a unique solution (i.e. no additional solutions within acceptable error limits), and always gives a best solution (i.e. one with the least cumulative error for all the site position estimates). This solution is often counter-intuitive. If using the Resolve program to map, it is best to work at a slightly higher error level for calculations of the initial double digest temporary solutions. This will tend to result in multiple rather than unique solutions for the temporary maps, but this is advisable as there is no guarantee that a partial solution found at an error setting of, say 2%, is necessarily correct relative to a solution found at, say, 3%, and is more likely to result in a minimum error final map in the 3-way analysis.

Lengthy perusal of an enigmatic or poor-quality autoradiograph in the hope of achieving sudden enlightenment is not advised. As time passes the banding patterns come to look more and more like a Rorschach Inkblot Test, with the accompanying transformation of mtDNA fragment analysis into self-analysis !

CHAPTER 7

METHODS OF INFERRING PHYLOGENIES

7.1. THEORY OF CLADISTICS

7.1.1 General Concepts

Measuring the ordered, transformative nature (and hence relationship) of characters over evolutionary time is the cornerstone of cladistic theory (Hennig, 1966). Cladistics uses analyses of character transformations to reconstruct evolutionary histories in the form of nodes (points designating common ancestry) and branches (which are pathways defining the evolutionary relationships between the ancestors).

A cladogram is a schematic representation of a reconstructed or inferred phylogeny (Li *et al*, 1991). The ordered, hierarchical patterns of radiation thus represented are based on the measured degree of similarity between homologous characters. Cladistic analysis does not use a set of empirically measured characters per se to define taxonomic relationships. Rather it is concerned with the pathways or ancestral lineages which such (transforming) homologous characters reflect (Eldridge *et al*, 1980) (homology meaning similarity as a consequence of common ancestry).

For example, maximum parsimony methods choose the cladogram which reflects the shortest pathway between the informative characters' transformative states as the closest reconstruction of the true phylogenetic tree (Li *et al*, 1991).

7.1.2 Homology

Cladistics uses shared derived characters (homologies) as informative sites. It is important to realise that cladistic character states such as homology, synapomorphy, symplesiomorphy and autapomorphy are relative designations made with specific reference to the phylogeny under study (Eldridge *et al*, 1980).

The concept "homology" is not new, but is inherent within Darwin's theory of evolution (Eldridge *et al*, 1980). A reformulated operational definition of the term was felt necessary so that such an important construct could be subjected to rigorous statistical analysis. In Eldridge's words (Eldridge *et al*, 1980, p.36), "homologous similarities are inferred inherited similarities that define sub-sets of organisms at some hierarchical level within a universal set of organisms".

Cladistic analysis rejects autapomorphic (that is unique) character states and characters that are similar throughout the group in phylogenetic reconstructions. For a character to be informative it must be present in at least two but in no more than n-2 taxa (where n = sample size) (Li *et al*, 1991).

The term homology is a relative concept with characters acquiring the said status with reference to the (hierarchical) level of a phylogenetic study (Eldridge *et al*, 1980). Hence, for example, a character which is common amongst all taxa at a low level of investigation (that is non-informative), can achieve homology (informative) status within a higher level, if it defines a sub-set within that hierarchy. Equally, a synapomorphy at the subfamily level may be transformed into an symplesiomorphy at a species level if a subsequent study at the species level only is undertaken.

7.1.3 The Cladistic Hypothesis

Eldridge *et al* (1980, p.50) defines the cladistic hypothesis as being "a cladogram specifying a pattern of relationships (nested sets) among taxa that is a consequence of a nested pattern of synapomorphy" - with synapomorphy being defined by Eldridge *et al* (p.53) as "the condition of sharing a derived character state or a later stage in the transformation sequence of a derived character state."

7.1.4 Character States

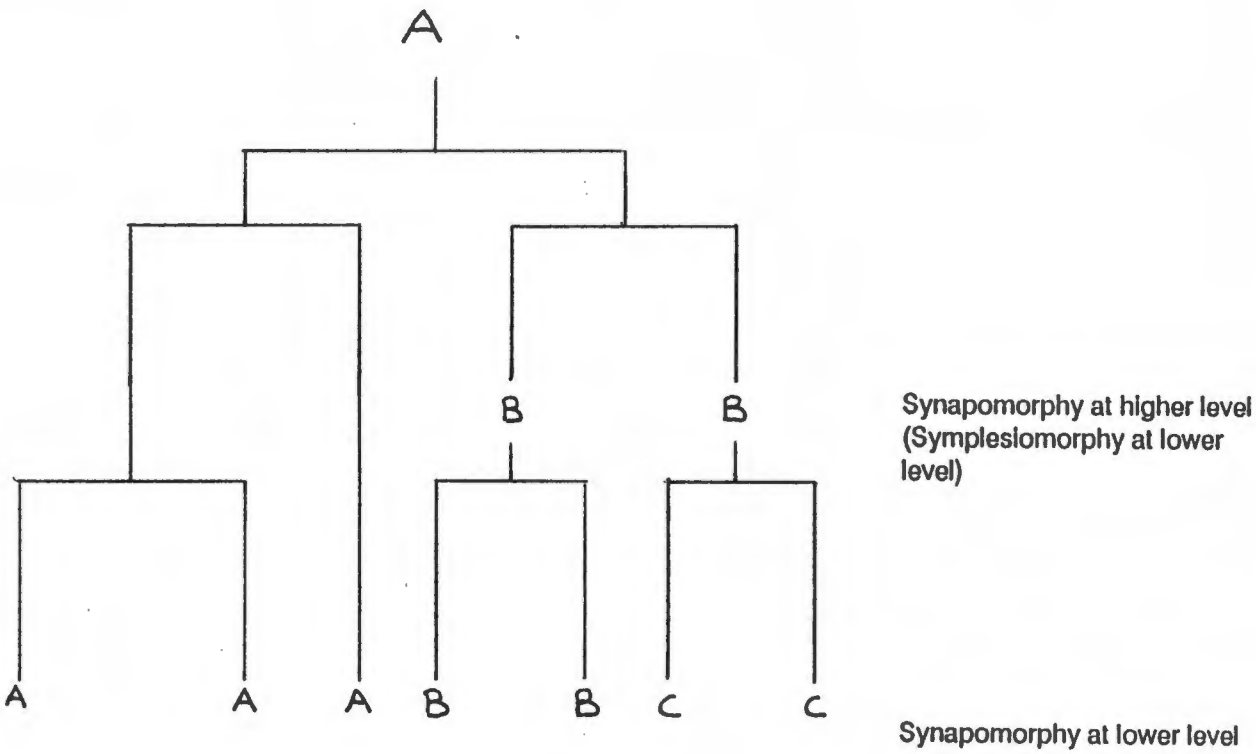
The key to a cladogram's construction appears to be the hierarchical separation of the shared similarities derived from an immediate common ancestor (synapomorphies) from those that are derived from an ancestor more remote than the immediate common ancestor (symplesiomorphies) (Eldridge *et al*, 1980).

The synapomorphic homologies are used to define the nested patterns at the low level of the phylogeny under study whereas the symplesiomorphic homologies are used to define the higher level nested patterns. It is evident that a symplesiomorphy (with reference to a given hierarchical level) cannot be used to define a sub-group at a lower level, as more general or distant characters cannot be used to group the more recent and hence more specific synapomorphic sub-groups (Eldridge *et al*, 1980).

Whether an homology assumes a synapomorphic or symplesiomorphic character status depends on the hierarchical level (with reference to the universal set) at which the study is undertaken. The character states are, as with the concept homology itself, a relatively defined condition. A synapomorphy at a higher level of phylogenetic analysis can become symplesiomorphic at a lower level.

(For example please see overleaf.)

Example of the relatively defined character states of homoplasies (symplesiomorphic or synapomorphic) :



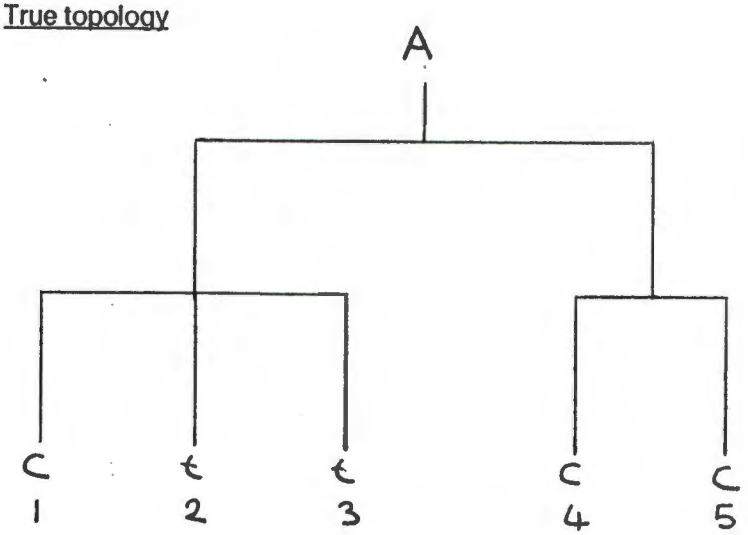
A = symplesiomorphy

7.1.5 Homoplasy as a Source of Error Variance

Separate mutational events or autapomorphies can lead to characters appearing to be similar in a group of taxa. They can therefore be taken (in error) as a shared derived character. Hence the term homoplasy denotes a cladistic error source. Parallel mutations or convergent evolution and back mutations are two sources of such an error.

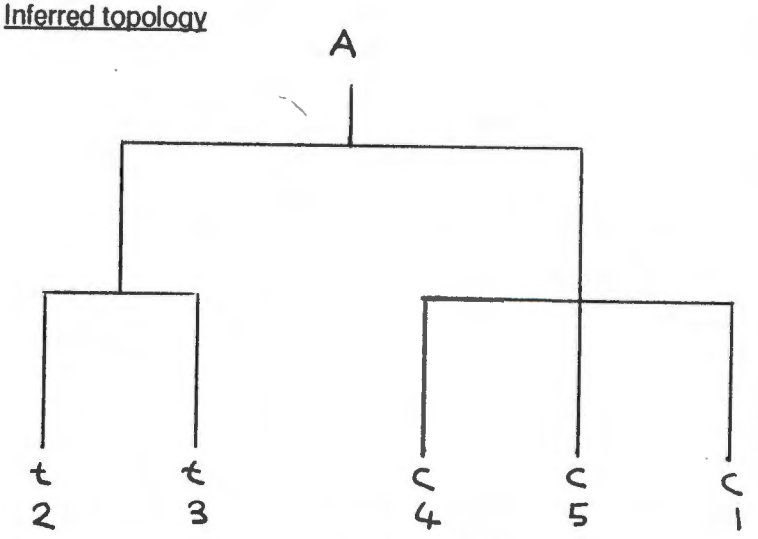
7.1.5.1 Example of a parallel mutation :

True topology



A is the shared primitive character (symplesiomorphy).
A mutates to t. Taxa 1, 2 and 3 should be grouped together on the basis of this synapomorphy. Taxon 1 subsequently undergoes an autapomorphic mutation (from t to c).
Taxa 4 and 5 are grouped together as they share the derived character c (A mutates to c).

Inferred topology

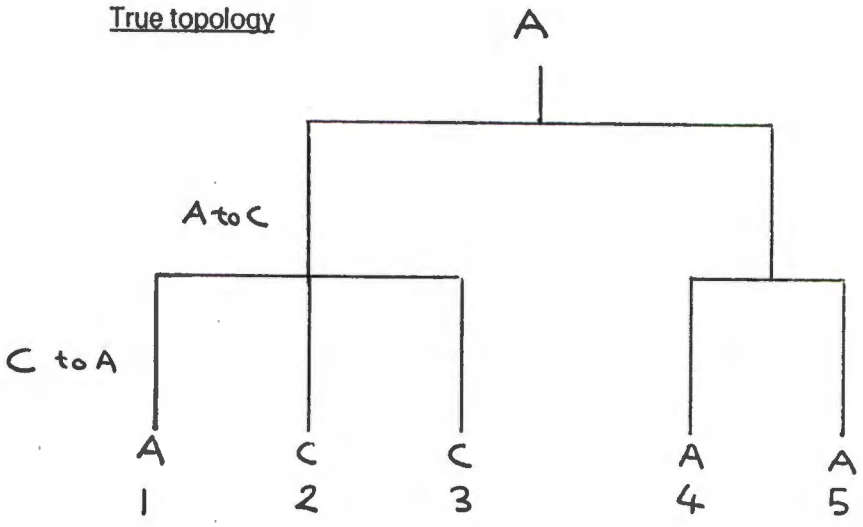


In the estimated topology taxon 1, on the basis of lower level autapomorphic mutation (t to c), will be falsely nested with taxa 4 and 5, which share a genuine synapomorphy.

7.1.5.2 Example of a back mutation :

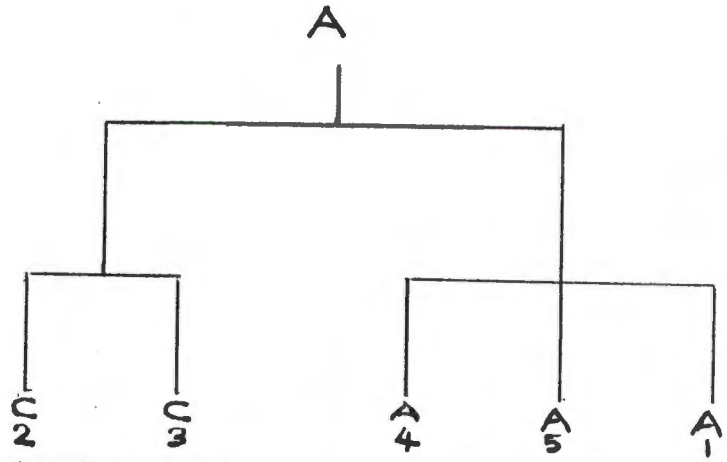
(A back mutation occurs when a character reverts or mutates back into its symplesiomorphic status.)

True topology



Character A is a symplesiomorphy and is an informative site clustering taxa 4 and 5 together.
A mutates into C, C being synapomorphic for taxa 1, 2 and 3.
Taxon 1 subsequently undergoes autapomorphic back mutation from C to A.

Inferred topology



In the estimated phylogeny taxon 1, because of its autapomorphic character transformation from C to A, will be (falsely) grouped with taxa 4 and 5 since they now all share the primitive character A.

7.1.6 A Cladistic Definition of Monophyly

A group of taxa is considered to be monophyletic if their reconstructed phylogeny estimates them to have radiated from a common stem. Cladistic theory therefore defines monophyletic groups in terms of synapomorphies. The immediate common ancestor shared by a group of taxa constitutes the stem of the monophyly (Eldridge *et al*, 1980).

7.2 PHENETICS OR DISTANCE MATRIX METHODS

Phenetics is the study of relationships among taxa on the basis of the degree of similarity between them (Li *et al*, 1991).

A pairwise sequence divergence matrix is constructed for all possible pairs of taxa under study. The pairwise distance data (which is a measurement of the degree of sequence divergence between all pairs) is subsequently used to construct dendograms.

If it is assumed that a constant relationship exists between evolutionary distance and divergence time and hence that there is a constant rate of evolution among the different lineages under study, pairwise sequence divergence data can be used to infer phylogenies (Li *et al*, 1991).

7.2.1 Sources of Error

7.2.1.1 Autapomorphies as a source of error

The unequal rate of evolution which leads to the accumulation of autapomorphic characters amongst taxa is a serious source of error variance (Li *et al*, 1991).

This is more probable for morphological data than for sequence data, as in general mutations occur independently of selective pressure. This is true for non-coding regions and for neutral substitutions but not for others, because of non-viability of some zygotes for changes in coding regions. With reference to morphological data, for example, a taxon isolated in a new and different environment from its ancestor would probably undergo more phenotypical (adaptive) changes than those taxa that remain in their original and more stable environment. Given that all the operational taxonomic units (OTU's) stem from an immediate common ancestor, such autapomorphic character transformations would make the taxon and its sister group appear more divergent in time than is in fact the case.

7.2.1.2 Homoplasies

Distance measures are also subject to homoplasic error.

7.2.2 The Transformed Distance Method

This method, devised by Farris (1977), attempts to control the error variance caused by the unequal rate of evolution among the taxa under study. Such a method uses an outgroup comparison to correct the original pairwise distance matrix of such effects. As previously mentioned, an outgroup is not too distant and shares a common ancestor to the taxa under study. It is by an appropriate set of (preferably separate) criteria clearly more distantly related to the group under study and to any of the within group taxa one to another.

There are algorithms which reconstruct the most probable phylogeny using the distance values of the transformed pairwise data matrix (Li *et al*, 1991). The transformed distance method uses the following formula to correct for unequal rates of evolution :-

Given :

Pairwise distance matrix for taxa A, B and C, with D as an outgroup

$$d'ij = [(dij - diD - djD)/2] + d_D$$

Where :

d'ij	=	transformed distance
i	=	A, B or C
dij	=	distance between paired taxa
djD or diD	=	distance between paired taxon and outgroup
d_D	=	(dAD + dBD + dCD)/3

The term d_D is introduced so that all transformed distance values remain positive. For the general case of n OTU's (not including the outgroup), $d_D = d_{iD}/n$.

(Li *et al*, 1991 p.109).

7.2.3 The Unweighted Pair Group Method with Arithmetic Mean (UPGMA)

This is the simplest of the distance methods used in phylogeny reconstruction. This method can be used in conjunction with the transformed distance method (which corrects for unequal rates of evolution) or it can be used with the assumption that rates of evolution are constant amongst the different lineages of the taxa under study (Li *et al*, 1991).

The UPGMA method clusters pairs of taxa in an hierarchical order of pairwise degree of similarity. This is achieved by grouping the most similar pair of taxa under study together. This pair is now considered singular in the subsequent pairwise comparison. Such a pair is termed a composite operational taxonomic unit (COTU). The pairwise comparison continues (grouping the most similar pairs together) until only two OTU's remain (Li *et al*, 1991).

7.2.4 The Neighbor-Joining Method (Saitou and Nei, 1987)

Contrary to the UPGMA method, this method is not based on clusterings of taxa, but rather the dendogram is constructed by linking the least distant pair of nodes.

The common ancestral node of the two linked nodes is added to the tree, and the two linked nodes (terminal nodes) with their respective branches are removed. The ancestral node is thus converted into a terminal node on a tree of reduced size. The two least distant pairs of terminal nodes are again replaced by a single common ancestral node. The process is complete when a single branch linking two nodes remains (Hillis *et al*, 1989).

The neighbor-joining method does not assume a constant rate of evolution. However, this method does assume that these data come close to fitting an additive tree (i.e. all pairwise distances are equal to the sum of the branch lengths that connect the respective taxa). As such, the pairwise distances between nodes is adjusted on the basis of their average divergence from all other nodes. This has the effect of normalising the divergence of each taxon for its average clock rate (taken from Hillis *et al*, 1989, p.442).

In support of distance measures based on sequence divergence for inferring phylogenies, Takahata *et al* (1991) have demonstrated (by using formulas for the minimum and maximum values of the sampling variance), the (statistically) satisfactorily accurate estimates of the sampling variance of nucleotide diversity as used by distance methods.

7.3 ROOTED AND UNROOTED TREES: OUTGROUP COMPARISON

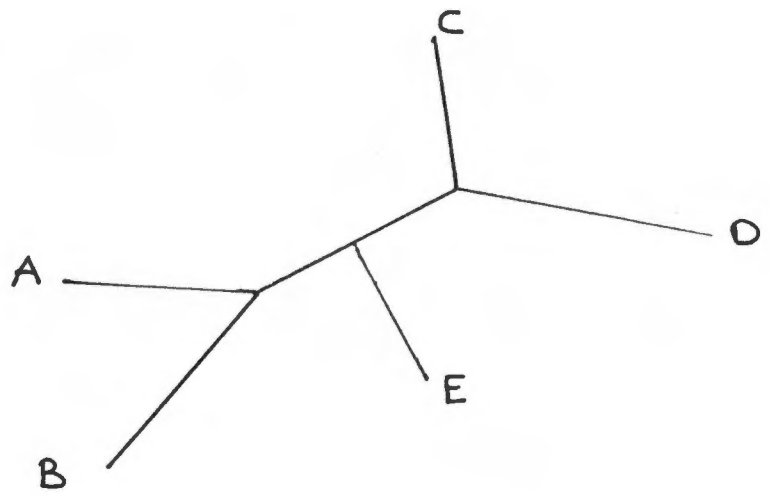
Phylogenetics is an attempt to reconstruct evolutionary histories. An unrooted tree is unable to infer such genealogical reconstructions as it only defines the relationships among the operational taxonomic units (OTU's) themselves, with there being no reference to ancestral lineage. A rooted tree's nodes and branches correspond to ancestral lineages, with each OTU's unique pathway or lineage being its inferred evolutionary history.

The tree's root is a known common ancestor of all the OTU's under study (Li *et al*, 1991).

The inclusion of a common ancestor "polarises" the relative set of relationships amongst the OTU's into an hierarchically ordered transformation of characters over evolutionary time (Eldridge *et al*, 1980).

(For diagrams please see overleaf.)

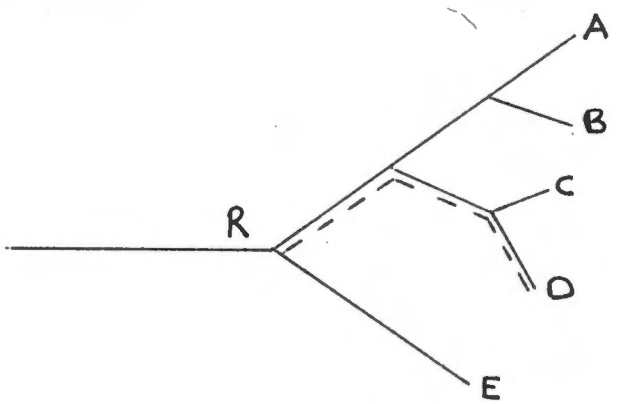
Diagram of an unrooted tree



Note that only the relationships amongst the OTU's themselves are specified.

(Li *et al*, 1991)

Diagram of a rooted tree



R = common ancestor

R to D = inferred unique evolutionary lineage of OTU "D"

Eldridge *et al* (1980, p.64) succinctly offer a cladistic interpretation of the use of an outgroup :-

> " The use of outgroup comparison is a search for the hierarchical level of distribution of each character state. Put another way, the method is a search for the level of the hierarchy at which each character state is a synapomorphy and therefore defines a set of taxa. At all lower levels that character is a symplesiomorphy."

The chosen outgroup taxon should share a relatively recent common ancestor to all the taxa under study. A taxon from a too distant outgroup will not have many shared primitive characters with the OTU's under study and hence will not give a statistically viable representation of the ancestral character states.

An outgroup taxon which is too closely related to the OTU's under study could cause an erroneous topology if in fact it is more closely related to some of the taxa than to others (i.e. it does not correctly represent the ancestral state).

Formally, an outgroup taxon should consist of those character states which define the larger monophyletic group of which the taxa under study are a sub-set. Commonly-shared characters between the outgroup and the taxa under study are considered to be primitive (plesiomorphic) whereas the characters that are found to be present only in the study group are hypothesised derivatives thereof (apomorphons) (Eldridge *et al*, 1980).

With reference to the earlier discussion on synapomorphies and symplesiomorphies, it will be understood that the designated primitive characters in any outgroup comparison are only relatively so, i.e. when compared with the relatively (hypothesised) more recently derived characters of the taxa under study. At a higher level of study the present outgroup's primitive characters could become synapomorphic.

CHAPTER 8

MOLECULAR EVOLUTION: RATE OF NUCLEOTIDE SUBSTITUTION

8.1 THEORY OF THE MOLECULAR CLOCK

Zuckerkandl and Paulings' (1962) initial observation of the apparently constant rate of amino acid substitutions in the haemoglobin and cytochrome c proteins among various mammalian lineages led them to propose the possible calibration of a "molecular clock" for any given protein. The translation of the rate of substitution into time, using a palaentological or comparative molecular data based method of calibration, constitutes the basis of the molecular clock theory. Important evolutionary events such as the time of divergence of two species can be estimated if the rate of nucleotide substitution is known and can be calibrated.

8.2 CRITICAL DISCUSSION OF THE MOLECULAR CLOCK HYPOTHESIS

When evaluating the validity of the molecular clock the following factors should be considered :-

(a) Variation in base substitution rate among lineages.

(b) Variation in base substitution rate among genomic regions.

(c) Accurate calibration of the clock.

(d) The degree of measurement error in sampling methods used.

(e) The stochastic nature of the rate of evolution and of mutational events.

8.2.1. Variation in Base Substitution Rate Among Lineages

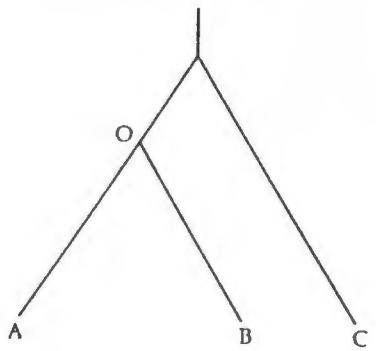
Relative-rate test

The translation of sequence divergence into time of species divergence has invoked great controversy (Hillis *et al*, 1990). In order to compare the rates of evolution between lineages independent of actual time, Sarich and Wilson (1973) constructed the relative-rate test.

In the comparison of the substitution rate of lineages A and B, the relative-rate test employs a third lineage which is known to have diverged earlier than the lineages A and B. This lineage C is then used as a reference to calculate the relative difference in the base substitution rate of lineages A and B.

(For Figure 11 please see overleaf.)

Figure 11 : Relative-rate test.



Where O denotes the common ancestor of lineages A and B

Kxy = is equal to the sum of substitutions that have occurred between any two given points.

It is evident from the illustrated phylogeny that :-

KAC = KOA & KOC
KBC = KOB & KOC
KAB = KOA & KOB

The values KAC, KBC and KAB can be directly estimated from the nucleotide sequences or indirectly from restriction site maps, the relative rates of substitution can be calculated by comparing the values from equations :-

KOA = (KAC + KAB - KBC) / 2
KOB = (KAB + KBC - KAC) / 2
KOC = (KAC = KBC - KAB) / 2

(Taken from Li *et al* (1990) pp 80 - 81.)

Using the relative-rate test for the comparison of the rates of synonymous substitution in mice and rats, Li *et al* (1987a) established the almost equal substitution rate between the two species. In a similar study using either the artiodactyl or carnivore lineage as a reference, Wu *et al* (1985) concluded that the synonymous substitution rate is about twice as high in rodents than it is in the human lineage.

Britten's (1986) analysis of sequence divergences from non-coding genomic regions led him to conclude that the rates of DNA change over different phylogenetic groups differ by a factor of up to five, with the slowest rates occurring in the higher primates and the fastest being observed in rodents, sea urchins and drosophila.

The proposed slower rate of molecular evolution in primates (hominoids in particular) when compared to other mammals, as suggested by Britten (1986), Li and Tanimura (1987) and Li *et al* (1987) has not been supported by Easteal's (1991) studies. In a comparative study involving 73 relative-rate tests, the latter researcher found only 1 significant ($P < 0,01$) difference in the evolutionary rate of 17 genes between humans and 6 nonhuman primate taxa.

The observed variance in substitution rates among lineages has been attributed to generation times. Lineages with similar generation times share a relatively constant evolutionary rate, whereas lineages with substantially differing generation times will have relatively different evolutionary rates (Li *et al*, 1987). Britten (1986) attributes this variation among lineages to evolutionary variation and selection of biochemical mechanisms such as DNA replication and repair.

8.3. BASE SUBSTITUTION RATE VARIATION AMONG GENOMIC REGIONS

Lake (1991) has noted that the unequal rate of evolution of sequence positions (as for example within the yeast rRNA) is a source of error in inferring phylogenies.

8.3.1 Base Substitution Rate

The rate of nucleotide substitution is defined as the number of substitutions per site per year. This rate can be calculated by dividing the number of substitutions between two homologous sequences by two (the time of divergence between the two sequences). When comparing substitution rates between genomic regions the same (taxa) pair should be used because -

(a) the rate of substitution may vary among lineages; and

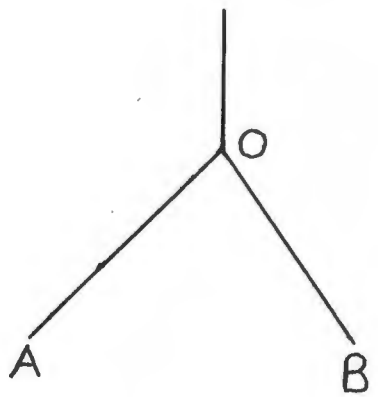
(b) palaentological calibrations are estimates of divergence times.

Restricting the comparison to the same pair will enable a relative estimate of substitution to be made, independent of a time frame.

8.3.2 Sequence Divergence

(For Figure 12 please see overleaf.)

Figure 12 : Calculation of sequence divergence.



Where A + B = extant taxa
O = last common ancestor
The sequence divergence between the two taxa is measured by summing the degree of divergence of both A and B from their last common ancestor;
or: the sequence divergence between A + B = AO + OB

The sequence divergence between any two taxa will be twice the base substitution rate x the time since divergence from the common ancestor, e.g. with a base substitution rate of 1%/million years, two taxa diverging from a common ancestor 2 million years ago would be expected to have a sequence divergence of 4%.

8.3.3 Rates of Base Substitution in nDNA and Coding and Non-coding Regions

8.3.3.1 Coding regions

The rate of non-synonymous substitution is very variable among genes from different lineages with a range from essentially zero in Histones 3 and 4 to $2{,}80 \times 10^{-9}$ substitutions per site per year in interferon γ.

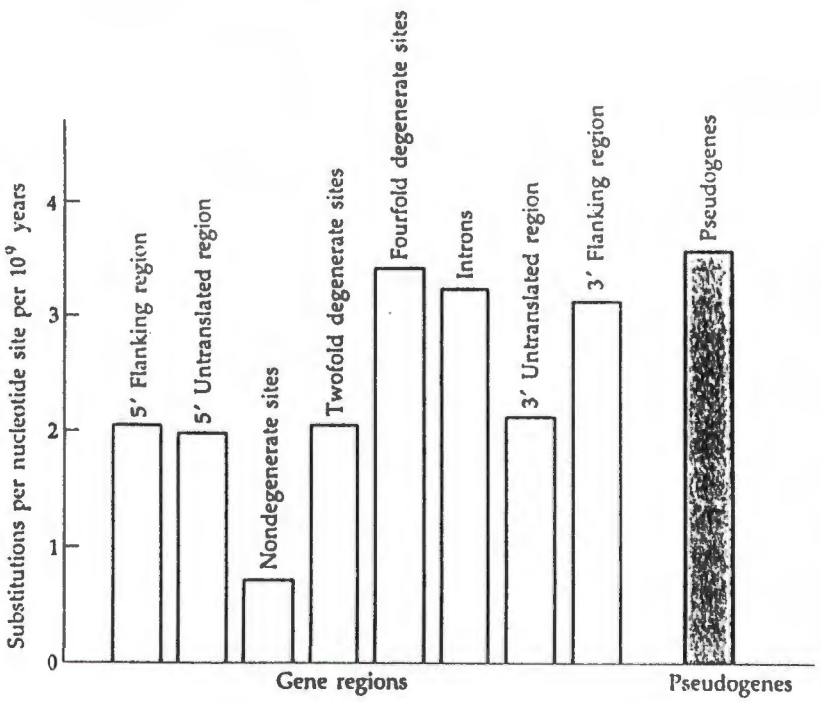
The rate of non-synonymous substitution appears to be determined by the selection intensity which in turn is determined by functional constraints.

The synonymous substitution rate also varies considerably, but less so than the non-synonymous rate. The mean rate of synonymous substitution (4,6 x 10^{-9} substitutions per site per year) is five times higher than the non-synonymous mean rate. This much higher rate of synonymous substitution is explained by the fact that synonymous changes will be selectively neutral since they cause no alteration in the phenotype and thus are more likely to become fixed in the population.

8.3.3.2 Non-coding regions

Substitution rates vary greatly among the 5' and 3' untranslated regions of transcribed genes, but this variation may largely represent sampling error since these regions are very short. Pseudogenes (i.e. copies of recognisable genes that are frequently found in a nonfunctional non-expressed state scattered throughout the genome) have the highest substitution rate of all non-coding regions (Li *et al*, 1990).

Figure 13 : Diagram of average rates of substitution in different parts of genes and in pseudogenes (taken from Li *et al*, 1990, p.73)



8.3.3.3 Mammalian mtDNA

The substitution rate in mtDNA for both synonymous and non-synonymous regions is much higher than that of nDNA. The non-synonymous rate varies greatly among the thirteen protein coding genes, but the synonymous rate is more constant and has been estimated to be $5,7 \times 10^{-8}$ substitutions per synonymous site per year (a value ten times higher than similar substitutions in nuclear protein coding genes (Brown *et al*, 1982).

8.4. CALIBRATING THE MOLECULAR CLOCK

Carlson *et al* (1978) are of the opinion that the calibration of the molecular clock constitutes a large error source since it is difficult to pin-point the last common ancestor of a group of extant taxa, even when using comprehensive fossil records. With reference to cetacean fossil material, this is a particularly pertinent point since fossil lineages of extant families are incomplete (refer Chapter 3).

Using a second type of molecular data such as protein sequences to calibrate the clock is not appropriate since errors associated with the calibration of the two clocks are compounded. In addition, this method assumes that the initial calibration (based on another group) is valid for the group in question (Hillis *et al*, 1990).

8.5. SAMPLING ERROR

The degree of sampling error incurred in the measurement of sequence divergence is a function of the resolution of the method used and the sample size. Sequencing, for example, is an extremely powerful technique since it entails the direct measurement of the order of bases on the selected DNA region. Hence one might expect little measurement error when using this method. These errors can be negligible until sufficient divergencies result in difficulties in ensuring correct alignment of homologous regions, a feature compounded as gaps or duplicate regions accumulate.

Restriction site mapping (RSM) differs in that it is an indirect sampling method (the RE recognition sequence positions on the mtDNA genome must be calculated, a method that incurs at least some degree of error. For example, it is uncertain as to whether sites mapped to within an acceptable error limit are truly homologous, as opposed to being two very close but non-homologous sites. It samples the whole mitochondrial genome, therefore the average rate of base substitution is used rather than the specific rate of substitution for the selected region as used by sequencing.

Sample size - Increasing the sample size will reduce the sampling error and is affected by either -

(a) increasing the size of the chosen DNA segment(s) to be sequenced; or

(b) increasing the number of cleavage sites in restriction site mapping. The number of correctly allocated synapomorphies should cancel out the misinformation produced by the (random) homoplasies in cladistic analysis. The stochastic error will be decreased in proportion to the increasing of data in distance estimates.

8.6. THE STOCHASTIC NATURE OF THE RATE OF EVOLUTION AND OF MUTATIONAL EVENTS

Goodman (1981) has noted the irregular rate of evolution within a population. He uses the example of the high rates of amino acid substitution which occurred following the gene duplication separating α and β haemoglobins and suggested that such high rates were due to advantageous mutations that improved the function of haemoglobin.

From this he contends that the rate of evolution often accelerates after gene duplication, and that protein sequences evolve much more rapidly at times of adaptive radiation.

Similarly, the stochastic model of genetic changes in populations assumes that changes in allele frequencies are not constant but variable, and can be predicted with only a degree of certainty.

There are a number of factors which determine whether a mutant allele will increase in frequency and eventually become fixed in a population. Examples of such factors are natural selection, random genetic drift, recombination and migration. In short, the change in allele frequencies with time is an irregular rather than a constant process.

8.6.1 Stochasticity of Mutational Events

A major source of error affecting sequence divergence comparison methods is the stochastic nature of mutational events. Base substitution is an irregular or random process (similar to radioactive decay), such that the correlation between mutation rate and time is not constant, but rather irregular. Erroneous estimates of divergence times can be deduced from sequence divergent data since the degree of sequence divergence is a function of the base substitution rate, which itself is irregular. Stochastic error is best controlled by increasing the sample size, since the stochastic error is inversely proportional to the amount of data accumulated. For example, in DNA sequencing it would be preferable to compare 1000 bp of sequence data rather than 100 bp.

8.7 SYNOPSIS

From the above discussion it is evident that :-

8.7.1 There is variation in the rate of base substitution among lineages.

8.7.2 There is variation in the rate of base substitution among genomic regions.

8.7.3 It is difficult to calibrate a molecular clock accurately.

8.7.4 There will be some degree of measurement error due to (a) sampling method; and (b) sample size.

8.7.5 Mutational events occur stochastically.

8.8 CONSTRUCTION OF A FEASIBLE MOLECULAR CLOCK

Molecular clocks are generally calibrated by dividing the average estimate of the age of the last common ancestor by the average measure of molecular divergence (Hillis *et al*, 1990).

8.8.1 Local Molecular Clock

8.8.1.1 A local molecular clock should be specifically calibrated for any given group of taxa or lineage under study since the base substitution rate can vary significantly between lineages, but probably only if they are quite widely divergent (i.e. at about family level).

8.8.1.2 The confidence limits of the palaeontologically based calibration of the local molecular clock should be specified, since the accuracy of calibration is a function of the quality of fossil records.

8.8.1.3 The same genomic region should be used in the sequence divergence comparison. The Neutral Theory of Molecular Evolution (Kimura 1968, 1983) hypothesises that genetic polymorphism and most evolutionary transformation is initiated through neutral mutations. The sequence divergence comparison of neutral changes is probably preferable to the comparison of changes which alter the genetic code. This is because such neutral changes are far less likely to have any phenotypical effect on the organism. Thus the rate of neutral base substitutions occurs independently of the phenotypical evolution of the organism and as such should not vary with changing selective pressures (Kimura, 1983).

Further, as the non-coding substitution rate is much higher than the coding regions' substitution rate, the sequence divergence comparison of non-coding genomic regions such as introns, spacer regions or pseudogenes, is a more powerful or sensitive method of DNA characterisation (non-coding DNA regions have high levels of polymorphism).

In the case of endonuclease restriction site mapping data, the average rate of substitution over the complete mtDNA genome should be used since this technique samples the whole genome. Nei *et al* (1979) have developed various statistical formulas to be used for estimating the number of nucleotide substitutions between two populations or species, based on the evolutionary change of restriction sites in mtDNA.

8.8.1.4 The stochasticity of mutational events can be partially controlled by increasing the sampling size, but must be taken into account when interpreting sequence divergence data. Because of the irregular nature of mutational events confidence levels of the regression line expressing the relationships between sequence divergence and time are usually large. (For calculation of confidence levels see Hillis *et al*, 1990, pp. 508 - 514.)

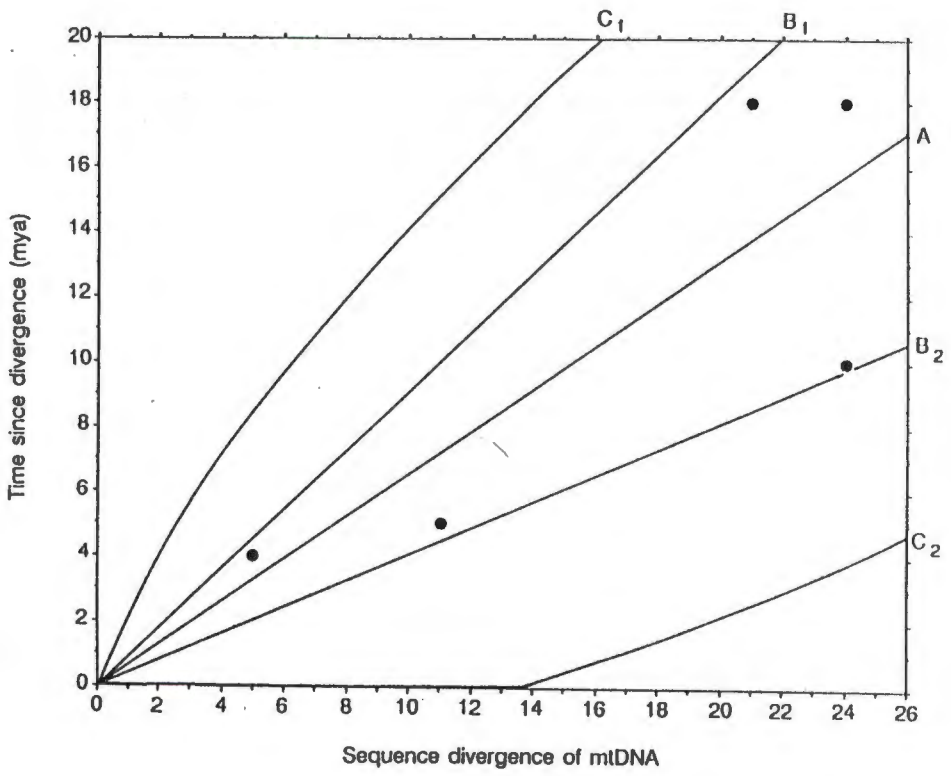
8.9. CONCLUSION

Although a morphological clock certainly does not exist (and hence the discredited nature of phenetic methods applied to morphological data) a fair case for a reasonably accurate molecular clock can be made, provided that - (a) stochastic error is reduced by a suitably large quantity of data, (b) neutral changes are studied, (c) homologous regions are compared, (d) comparisons are restricted to reasonably close lineages, and (e) absolute time calibrations are used with caution (relative time scales are equivalent to sequence divergence).

8.10 EXAMPLE OF THE CORRELATION BETWEEN SEQUENCE DIVERGENCE AND TIME

In Figure 14 mtDNA sequence divergence of primates is correlated with time using primate fossil records to calibrate the relationship (figure from Hillis *et al*, p.512, 1990).

Figure 14 : Sequence divergence of primates correlated with time using primate fossil records to calibrate the relationship



Key : Regression (A) of estimated time since separation on sequence divergence of mt.DNA in primates.
B_1 and B_2 are the bounds of the 95% confidence limit of the regression line.
C_1 and C_2 are the bounds of the 95% confidence limits for predicted values of time based on new measurements of sequence divergence, except that regions of negative time are collapsed to zero (Hillis, 1990, p.512).

CHAPTER 9

METHODOLOGY

9.1 SOURCES OF BIOLOGICAL MATERIAL

Heart tissue from all eleven species examined was obtained from strandings or from animals that drowned accidentally in commercial trawl nets. The only exception was a specimen of *C. heavisidii* which was taken under permit for P B Best's research, of which the heart was donated for use in the present study.

The heart muscle was removed as soon as possible after death and stored frozen (- 20 °C) until use. The skeletal material from the specimens used is in the possession of the S A Museum. All strandings were collected along the South African coastline within the Cape Province boundary.

(For Table 1 please see overleaf.)

TABLE I

SOURCES OF BIOLOGICAL MATERIAL

TAXON	ACCESSION NO.	SITE OF STRANDING
Mesoplodon layardii	ZM.40857	St Helena Bay Harbour
Hyperoodon planifrons	ZM.40855	Berg River mouth, St Helena Bay
Globicephala melas	(P.B. Best 91/10)	Die Plaat, Hermanus
Feresa attenuata	ZM.40867	Dwarskersbos, St Helena Bay
Grampus griseus	ZM.40860	Buffels Bay, Cape Point Reserve
Cephalorhynchus heavisidii	ZM.40900	Taken at sea 32 ° 25,4'S, 18 ° 14,6'E
Delphinus delphis	(P.B. Best 91/04)	Arniston
Lagenorhynchus obscurus	ZM.40911	Died in commercial trawl net 19 ° 19,7'S, 12 ° 32,1'E
Tursiops truncatus	(P.B. Best 90/36)	Mossel Bay
Stenella coeruleoalba	ZM.40919	Skipskop, Cape Agulhas
Caperea marginata	(P.B. Best 90/12)	Simonstown

9.2 METHODOLOGY

9.2.1 Protocol for Mitochondrial DNA Extraction

Mitochondrial DNA isolation using the Cesium Chloride - Ethidium Bromide (CsCl-EB) Gradient Protocol

The protocol is divided into two main phases. The first entails the preparation of a mitochondrially-enriched fraction which is achieved through a series of differential centrifugations of the tissue homogenate. The second stage entails the isolation of the mitochondrial DNA (mtDNA) from the nuclear DNA (nDNA). The two DNA's are separated on the basis of their differing conformations. The linear nDNA molecules tend to bind more with the intercalating dye ethidium bromide (EtBr) than do the supercoiled mtDNA molecules. The intercalating dye, being of a lower density than DNA, decreases the density of the nDNA, thereby increasing the relative difference in densities between the mtDNA and the nDNA. Subsequent centrifugation separates the two DNA's with the (lighter) nDNA band forming above the mtDNA band in the appropriately created CsCl gradient (Ausubel *et al*, 1991) (Sambrook *et al*, 1989). The lower mtDNA band is removed and purified.

9.2.2 Preparation of crude Mitochondrial DNA

9.2.2.1 Wash phase

An 80 gram sample of frozen heart tissue is shaven into thin sections using a scalpel blade. Care must be taken to remove any fatty tissue. For practical purposes the sample is divided into two 40 gram units and resuspended in 4,5 x volume of extraction buffer (100 mM Tris-HCL, pH 7,4; 150 mM NaCl; 20 mM EDTA, 10% (w/v) sucrose). The solutions are each transferred into a 200 ml polypropylene screw-cap JA 14 tube and centrifuged at 1000 g for 10 minutes at 4 °C in a Beckman VI 60 centrifuge. The wash phase cleanses the heart tissue of any contaminants that may have occurred during retrieval and storage of the material as well as beginning the equilibration of the tissue with a solution of appropriate pH, salt and EDTA concentrations.

9.2.2.2 Homogenising phase

The supernatants from the wash phase are decanted and discarded. The pellets are resuspended in 3 x volume of extraction buffer and homogenised for 25 seconds in a Waring Blender at full speed.

The two samples are homogenised separately to ensure a thorough breakdown of the tissue. The homogenate is transferred into 200 ml JA 14 tubes and centrifuged at 1000 g for 10 minutes at 4 °C.

9.2.2.3 Resuspension in STE

The supernatant in which the mitochondria and other small cellular debris is suspended is filtered through cheesecloth to remove residual fat particles. The pellet consisting of nuclei and large cellular debris is discarded. The supernatant from both tubes is transferred to a single 200 ml JA 14 tube and centrifuged at 10000 g for 15 minutes at 4 °C, to pellet the mitochondria. The supernatants of both tubes are combined at this stage as there is less wastage involved in the retrieval of a single large pellet than that from two smaller pellets. The supernatant is decanted and discarded, and the pellet is drained. Using a pipette the pellet is resuspended in 4 ml STE buffer (100 mM NaCl; 10 mM Tris-HCl, pH8,0; 1 mM EDTA). The tube is rinsed thoroughly.

All manipulations up to this stage are performed at 4 °C. All equipment such as rotor heads, centrifuge tubes and the Waring blender must be pre-cooled to 4 °C before use. After this stage, work at room temperature.)

9.2.2.4 Lysis phase

The purpose of this phase is to lyse the mitochondrial membranes. Remeasure the volume of the suspended pellet. Add sodium dodecyl sulphate (SDS) to 1% and mix gently. The solution goes from opaque to semi-clear. Allow to stand for 20 minutes.

9.2.2.5 Precipitation of excess proteins

Add cesium chloride (CsCl) to 1 M. CsCl has a molecular weight of 168,36 $gmol^{-1}$. Therefore use 168,36 g for 1000 ml of solution or for *z* amount of solution use -

$$\frac{168.36 \times z \text{ g CsCl}}{1000}$$

Allow to stand for at least 30 minutes to facilitate precipitation of excess proteins. Transfer the resuspended pellet into a 50 ml polypropylene screw-cap centrifuge tube and centrifuge at 1000 g for 10 minutes at 20 °C. This operation is performed on a Sigma 2 MK machine using the 1200 V rotor at 10000 rpm. Decant the supernatant into a 20 ml culture tube. Discard the pellet.

9.2.2.6 Separation of the mtDNA / nDNA phase

Ethidium bromide (EtBr) is a low density intercalating dye which bonds more readily with the linear nDNA than it does with supercoiled mtDNA.

The volume of solution is remeasured. An 80 ul stock solution (10 mg/ml) ethidium bromide per millilitre of solution is added. Gloves are worn when using EtBr since it is carcinogenic. EtBr stock solution is wrapped in foil as it is ultra violet (UV) light sensitive.

Creation of the Cesium Chloride (CsCl) density gradient phase :

The ideal density of the solution is 1,55 gms/ml. To attain this density add 1 gram of CsCl for each millilitre of solution, minus the amount added in the precipitation phase. For example :

Volume of solution = 4,5 ml
Add 4,5 gms CsCl - 0,9 gms CsCl (amount used in precipitation phase)
= 3,6 gms CsCl to be added

Before measuring the density of the solution, calibrate the 1000 ml pipette (1 ml distilled water = 1 gram). Once the adjustable pipette is calibrated draw 1 ml of the solution and weigh. Add CsCl to increase or STE buffer to decrease the density, until the correct density of 1,55 g/ml is attained. Allow the CsCl to dissolve fully before measuring or remeasuring the density. A density of 1,55 g/ml is found to be ideal as the DNA bands should stabilise as its buoyant density about one-third from the top of the ultracentrifuge tube after centrifugation.

Too high a density will compact the two bands at the top of the tube, whereas low-density will cause both bands to stabilise at the base of the tube, causing the mtDNA to become contaminated with the nDNA. The solution is transferred, using a 5 ml syringe, to the 5 ml Beckman heat sealable tubes. Fill the tube to the base of the neck. Press the needle against the side of the tube to avoid bubbles. Dry the inside of the neck to ensure proper heat sealing. Centrifuge at 50000 rpm at 20 °C for 16 to 20 hours in a Beckman VI 60 Vertical Rotor.

9.2.2.7 Recovery of the mtDNA

Remove the ultracentrifuge tubes carefully from the rotor. In natural light the linear DNA (mostly nDNA with some relaxed circular DNA or damaged mtDNA) should be visible as an intense red band. The mtDNA band, which probably won't be visible, should be between 4 mm - 7 mm below the nDNA band. Mark this region with a felt-tip pen in case the mtDNA band is not visible even under UV light. Gently clamp the tube on a test-tube stand and sever the neck with a scalpel blade. From now use a long-wave (305 nm) UV light source in an otherwise dark room. Wear safety glasses and gloves. Try and locate the mtDNA, which should be visible as a faint orange band. If it is not located, remove the solution from the area previously marked. This is achieved by using a peristaltic pump.

Flush the pump alternatively with ethanol and distilled water before and after use. Place the drain pipe into a 20 ml sterident vial. Before lowering the suction tube through the nDNA band create a bubble on the tip by gently pumping some air through. This will prevent contamination as the tube passes through the nDNA band. Slowly transfer the mtDNA band region from the ultra-centrifuge tube into a sterident vial.

9.2.3 Final Purification of the mtDNA

9.2.3.1 Removal of the Ethidium bromide

Measure the volume of the solution. Add an equal volume of isoamylalcohol (equilibrated with CsCl - saturated water; the top layer is the isoamylalcohol). Shake the solution. The EtBr partitions into in the organic phase, whereas the mtDNA remains in the aqueous phase. Therefore, using a 1 ml pipette, remove and discard the top pink organic phase. Repeat this process a few times until the aqueous phase is completely clear. Remeasure the volume of the solution and transfer into 10 ml centrifuge tubes.

9.2.3.2 Dialysis

Using a 1 ml pipette add 2 x the solution volume of distilled water followed by 6 x the original solution volume of 97% ethanol, and mix thoroughly.

Store the solution for 1 to 2 hours at -20 °C to facilitate the precipitation of the mtDNA. Centrifuge at 10000 g for 10 minutes at 4 °C. Carefully discard the supernatant by pouring off away from the side of the pellet. (The pellet will form on the lower outer region of the tube after centrifuging.)

9.2.3.3 Wash phase

This phase purifies the mtDNA of any remaining cesium chloride. Resuspend the pellet in 5 ml 70% ethanol (EtOH). Vortex briefly, then centrifuge the solution at 1000 g for 10 minutes at 4 °C. Again, carefully decant and discard the supernatant by pouring away from the side of the pellet. Remove the screw cap from the centrifuge tube and place a perforated plastic sheath over the neck. Place the sample in a vacuum drying chamber for 5 to 10 minutes to remove the remaining water and ethanol from the pellet.

9.2.3.4 Storage phase

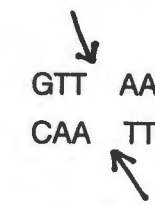
For long-term storage the mtDNA pellet is resuspended in TE buffer (see Appendix I). Resuspend pellet in 500 uTE buffer. Vortex thoroughly for 1 to 2 minutes, making sure to concentrate the solution on the (previously marked) pellet side of the tube. To prevent contamination aliquot into five 100 l samples (store at -20 °C.

9.3 DIGESTION OF mtDNA WITH RESTRICTION ENDONUCLEASES

9.3.1 Restriction Endonucleases

Nucleases are enzymes that cleave the phosphodiester bonds of polynucleotide chains. Endonculeases are those nucleases that preferentially cut internal bonds with exonucleases cleaving the terminal nucleotides. Restriction enzymes (REs) are nucleases that cleave the DNA double helix at a very specific nucleotide sequence, and the first one was isolated by Hamilton Smith in 1970 (Watson, *et al*, 1987). REs cleave both strands of the DNA since their recognition sequences are in most cases rotationally symmetrical around the centre.

For example, RE Hpa I's restriction site is :



with the arrows marking the point of cleavage.

REs are found only in procaryotic organisms, with their natural function probably being the protection of the bacteria from foreign DNA found in such bacteriophages. Over 400 RE have been isolated and characterised (Roberts, 1984). Isoschizomers are those REs that have been isolated from different types of bacteria, but which share the same recognition sequence. In general there is a preponderance of guanine and cytosine bases in the recognition sites (Hillis *et al*, 1990).

The REs cleavage sites can either be blunt-ended (the two strands are cut on centre) or staggered (the two strands are cut off-centre resulting in a short single stranded tail at the cleavage site - such a staggered cut can produce ends with with either a 5' or a 3' overhang).

REs generally recognise sequences of 4, 5, 6 and 8 nucleotides in length. The present study utilised only those REs that recognise 6-bp sequences, since these cleave the mtDNA into a manageable number of fragments for subsequent computational analysis. REs that recognise 4-bp sequences recognise about sixteen times as many restriction sites on the mtDNA genome. Double digests composed of such multi-cutting enzymes are proportionately more difficult to resolve as the number of possible fragment combinations is that much greater. The 6-bp REs used in this study variously recognise between zero and nine sites.

9.3.2 Protocol for the digestion of mtDNA with restriction enzymes

Storage : Enzymes are stored in 50% glycerol to prevent denaturation by freezing. The glycerol can affect enzyme activity if it is at greater than 5% present in the final reaction volume. REs are best stored at - 20 °C in a freezer. When in use keep on ice, and even then for as short a time as possible.

Use : REs vary in stability, with those that denature rapidly best used in greater quantities than those which are more stable.

REs function optimally under specific conditions with variables such as temperature, pH and salt concentrations affecting their activity. In this study conditions for RE digestion followed the specification of the supplier (Amersham International, Boehringer Mannheim, New England Biolabs) if the more frequently used 1 x or 2 x KGB buffer (McLelland *et al*, 1988) proved ineffective.

9.3.2.1 Single digest of a single concentration

At least 1,5 ng DNA is required for effective end-labelling (Hillis *et al*, 1990). In the present study approximately 1/400 th of the total yield of mtDNA per taxa was used in each digest. The amount of mtDNA needed per reaction was based on the quality of earlier autoradiographs. The final volume of the reaction (including the loading buffer) should not exceed the volume of the well in the gel (i.e. 15 - 20 ul for a 15 lane, 150 ml agarose gel). Water (distilled, de-ionised and sterilized) is added to bring the reaction to volume.

Example :

approx. 1 ul mtDNA(depending on stock concentration)
units restriction enzyme (as specified by manufacturer)
ul buffer (as specified by manufacturer)
6 ul loading buffer (refer Appendix IV)
ul distilled water (to volume)

Mix thoroughly by centrifuging briefly. Incubate at 37 °C. Depending on the amount of enxyme used, digestion of pure mtDNA is completed within 2 to 3 hours. After digestion store at -20 °C until needed.

9.3.2.2 Double digest of a single sample

The protocol for a double digest reaction is similar to that of a single digest except that the calculated units of the two chosen REs (that constitute the double digest) are added simultaneously to the reaction, and a general buffer (e.g. KGB x 1 or x 2) is used.

9.3.2.3 Multiple samples of single and double digests

Usually a carefully constructed order of a number of single and double digest reactions are undertaken simultaneously to enable direct comparison of their molecular weights and their banding patterns. A comprehensive discussion of the theory of double digests and restriction site mapping is undertaken in Chapter 5. When working with multiple samples the variables to be considered are the number of samples (usually 14 plus 1 marker on a 15 lane agarose gel) and the selection of single enzyme combinations for the desired double digest reactions. The multiple sample protocol is otherwise similar to those of single and double digest one-sample reactions. It is useful for those samples that are to be digested with the same REs to prepare a "cocktail" or "digest mix", then add an aliquot to each sample.

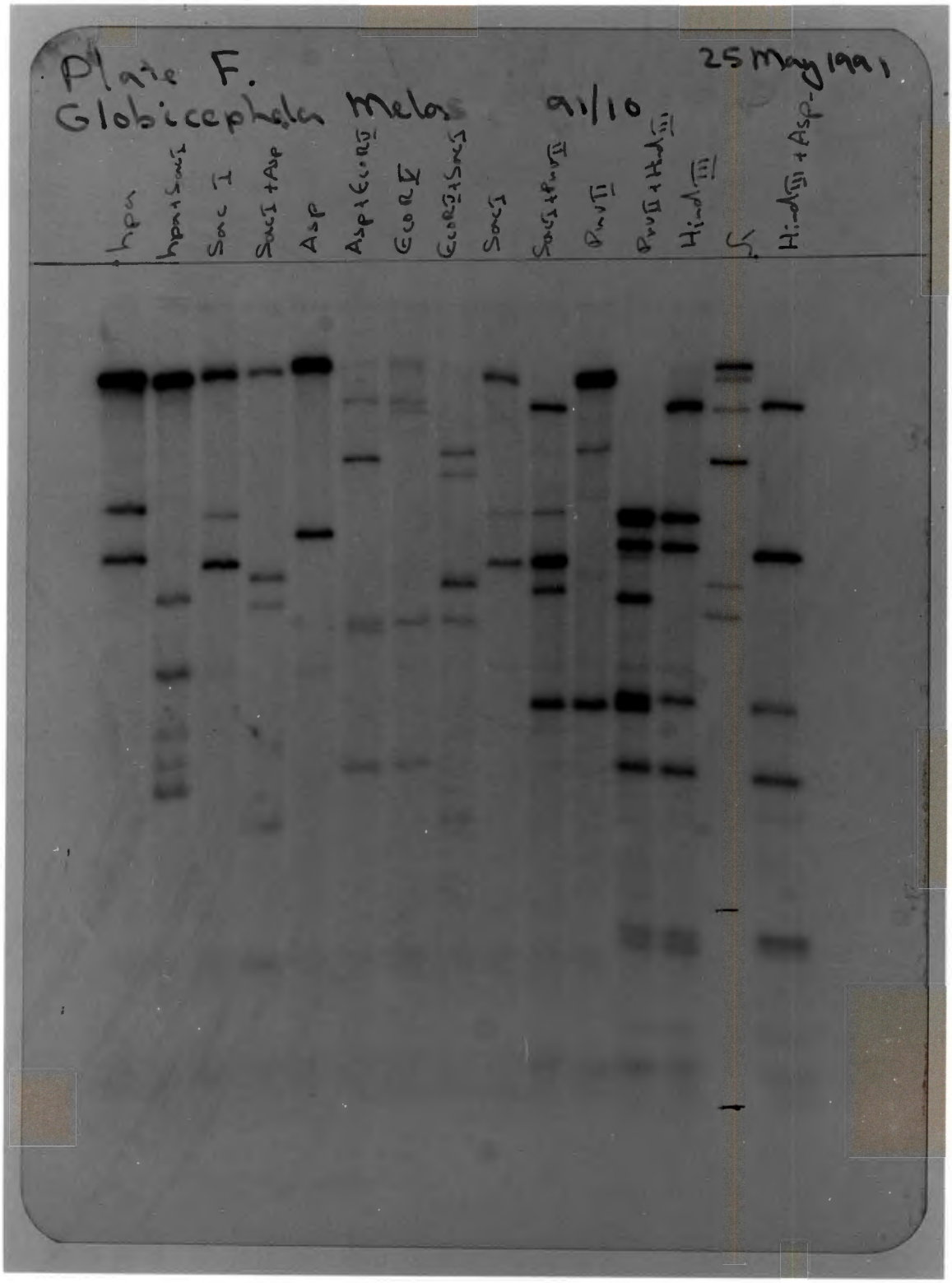


Figure 15 : An autoradiograph showing multiple samples of single and double digests (*Globicephala melas* mtDNA). The molecular weights (mw's) of the fragments are calculated using the known mw of phage Lambda's (cleaved with Hind III) fragments. (See Appendix II for Lambda's mw's.)

9.4 END-LABELLING WITH ^{32}P

Visualisation of the RE cleaved mtDNA fragments is achieved by end-labelling them with [^{32}P] or [^{35}S] deoxynucleotide triphosphates (dNTP). DNA fragments can be end-labelled using any of the four dNTP's (deoxyadenosine triphosphate [dATP], deoxycytidine triphosphate [dCTP], deoxyguanosine triphosphate [dGTP] or deoxythymidine triphosphate [dTTP]), as long as their compliments occur on the 5' overhang. Hillis *et al* (1990) recommend using all four [^{32}P] or [^{35}S] dNTP's when using several different restriction enzymes.

The present study found it sufficient to use ^{32}P - dCTP (cytidine triphosphate radioactively labelled with phosphorus - 32 [a position]).

Intensity is independent of fragment size as each fragment has the same number of ends. Hence a 100 bp fragment should be visually as intense as a 10000 bp fragment. However, possible band spreading (fragment dissipation) of the lower mw fragments can cause the bands consisting of low mw fragments to be more faint on the autoradiograph. [^{32}P] dNTP's have a higher energy emission but a shorter half-life than [^{35}S] dNTP's (fourteen days as opposed to sixty days), ensuring relatively shorter exposure times for autoradiography. ^{32}P end-labelling is a very sensitive technique with cleaved fragments from 1 - 5 ng DNA being visualised (Hillis *et al*, 1990).

9.4.1 Endonuclease Phase

The 5'to 3' exonuclease function of the large fragment of *E. coli* DNA Polymerase 1 (Klenow Polymerase; see Appendix I) converts blunt ends or 3' overhangs to 5' overhangs in absence of nucleotide triphosphates, which expose sites complimentary to the labelled nucleotide. A working concentration of 1 unit enzyme/ul per reaction of Klenow polymerase is added to the reaction and pre-incubated for 10 to 20 minutes at RT°.

9.4.2 Polymerase Phase

Thereafter dATP, dGTP, dTTP (to a final concentration of 2 mM; see Appendix I) and [^{32}P] - labelled dCTP are added to the reaction and further incubated for 10 to 20 minutes at room RT°. [^{32}P] - dCTP has a stock concentration of 10 ci/ul. 1 ci/ul is needed per reaction. [^{32}P] dNTP's have a half-life of two weeks, thus for up to two weeks use 0,1 ci/ul per reaction, and for two to four weeks use 0,2 ci/ul. The 3' to 5' polymerase function of the Klenow fragment results in the radioactive end-labelling of individual mtDNA fragments.

9.5 AGAROSE GEL PREPARATION AND ELECTROPHORESIS

It has been demonstrated (Harley *et al*, 1973 b) that the degree of mobility of DNA molecules migrating under electric charge through agarose or polyacrylamide gels is a function of their conformation and molecular weight (mw).

A DNA molecule's mobility is, over a fairly wide size range, inversely proportional to its mw. Considering DNA molecules of the same mw, closed circular conformation (ccc) DNA migrates most rapidly, followed by linear and nicked open circular (noc) DNA molecules. MtDNA has a closed circular conformation (Watson *et al*, 1987). Restriction endonuclease activity cleaves this cccDNA into linear fragments of varying mw. MtDNA fragment analysis techniques are concerned with the separation and measurement of the cleaved, linear fragments.

High density agarose gels (approximately 2,4%) are used to detect the small mw fragments (100 to 300 bp's), whereas the larger fragments (approximately 10 k bp's) are more accurately measured on low density gels (0,6% to 1%). As further discussed under Restriction Enzyme (R.E.) map construction, it is preferable when analysing complex double digest banding patterns to electrophorese the fragments through both high and low density gels for optimal results.

All the digests in the present study were run on horizontal agarose gels. Although both high and low density gels were utilised, the optimum density agarose gel for cetacean material was found to be 1,6%.

9.5.1 Protocol for Agarose Gel Preparation

The standard mould used measures 100 mm x 150 mm, and has a volume of 150 to 200 ml. The generally high resolution banding patterns obtained made the 15 lane gel ideal for multiple and economical double digest fragment analysis.

9.5.1.1 Mixing the agarose

Prepare a 1200 ml solution of TAE buffer (24 ml 50 x TAE (refer Appendix I), 1176 ml distilled water) in a two-litre flask. Decant 150 ml for the making of the agarose gel, with the remaining 1050 ml being used to submerge the gel in the electrophoresis bath to prevent it from desiccating.

Decide on the gel density to be used.

Example: For a 1% 150 ml agarose solution use

1/100% x 150ml/1 = 1,5 gms agarose

Mix the ingredients thoroughly in a flask and boil vigorously, swirling the solution intermittently, or microwave on high for 3 to 4 minutes, using a teflon-coated stir bar to avoid superheating.

The preparation is ready when all the particles have gone into solution. Visually the uncooked agarose solution is a milky opaque colour, with the ready solution being quite clear. Check the final volume and add distilled water to original volume to compensate for any evaporation. Allow to cool at RT° to 45 to 50°C. A solution that is too hot will warp the perspex casting tray. Pour the solution evenly, removing any bubbles that might form. After pouring check the comb's pre-aligned position. Allow to set at RT° for two hours.

Carefully remove the comb, taking care not to tear the walls of the wells. Remove the tape and place the mould (containing the gel) into the bath of the levelled electrophoreses apparatus with the wells close to the cathode. Fill the bath with the remaining (1050 ml) TAE buffer, avoiding the accumulation of air bubbles beneath the mould.

9.5.1.2 Agarose gel drying

Remove the mould from the bath. Place a single sheet of Whatman 3MM chromatography paper on the exposed side of the gel. Turn the mould over and gently prize the gel off. Place a thin plastic sheet on the now exposed side of the gel. Centre the gel (with the paper side down) on two sheets of chromatography paper placed in the bed of the gel-dryer. Set the temperature at 50 to 60 °C for agarose gels and up to 70 °C for polyacrylamide gels. Ensure that the gel is vacuum sealed. Allow to dry for 90 minutes.

9.5.1.3 Electrophoresis

The DNA's negatively charged sugar phosphate backbone will cause it to migrate towards the anode. Fragments are best resolved using low voltages, with full length double digest gels being run overnight.

In the present study fragments were separated according to their respective molecular weights through electrophoresis for 16 hours at 35 Vcm^{-1}. Electrophoresis is stopped when the dye (equivalent to 500 bp's in a 1% agarose gel) has migrated to the three-quarter mark on the gel. This usually ensures visualisation of the 125 bp Lambda Hind III marker.

Caution : Electrophoresis should be carried out in a separate area to avoid radio-active contamination. When working with radio-active materials -

1. Work behind perspex shields.

2. Wear gloves.

3. Preferably use micro-pipettes and centrifuge machines that are designated for radio-active use only.

4. Treat the dried gels as being radio-active.

5. After electrophoresis the TAE buffer in the bath is radio-active since it now contains the unincorporated ^{32}P - dCTP's.

6. Handle all radio-active waste such as used autoradiographed gels, micro-pipette disposable tips, eppendorfs, TAE buffer and gloves, with the appropriate care, and ensure proper disposal.

7. Check working space and equipment used for radio-active contamination by sweeping with a Geiger counter.

9.5.1.4 Autoradiography

Using tape affix the dried gel (sandwiched between the plastic and chromatography paper) into the autoradiograph cassette. Working in a darkroom, place the film (Amersham Hyperfilm - MP X-ray film) directly above the gel and seal the cassette. Store at -70 °C for 10 to 50 hours before developing. This temperature increases the sensitivity of the banding patterns. Exposure times depended on the amount of mtDNA labelled, the effectiveness of the end-labelled reaction and the age of the gel.

The use of intensifying screens is optional. A single intensifying screen enhances the intensity of the image by a factor of 4, with two screens intensifying this factor to 10. Monitor the gel with a Geiger counter and utilise past experience to decide whether or not to use such screens, since they can reduce the crispness of the image. Most of the autoradiography done in the present study utilised a single intensifying screen. After exposure, autoradiographs were developed, fixed and dried.

CHAPTER 10

RESULTS

10.1 CETACEANS SAMPLED

10.1.1 Family Delphinidae :

Delphinus delphis
Tursiops truncatus
Stenella coeruleoalba
Lagenorhynchus obscurus
Cephalorhynchus heavisidii
Grampus griseus
Feresa attenuata
Globicephala melas

10.1.2 Family Ziphiidae :

Mesoplodon layardii
Hyperoodon planifrons

10.1.3 Family Neobalaenidae :

Caperea marginata

10.2 ENZYMES USED

TABLE II

ENZYME	CODE	ENZYME	CODE
Bcl	c	Hind III	H
Sac I	s	NCO I	N
Sac II	S	Asp 718	A
xba I	X	Dra I	D
Cla	C	xho I	o
Bam H1	B	Hpa I	h
Eco RV	R	Bst E11	t
Pst I	P		

10.3 OUTGROUP

The baleen whale *Capera marginata* (pygmy right whale) was used as an outgroup in order to provide a root for trees produced by cladistic analysis.

10.4 RESTRICTION SITE MAPS

The length of the mtDNA molecule was estimated (using the sum total of restricted mtDNA fragments of digests consisting of fragments smaller than 6000 bp long), to be about 16400 bp long.)* Figure 16 compares the restriction sites for fifteen enzymes recognising six base pair sequences for eight members of Delphinidae, two members of Ziphiidae and the only member of the family Neobalaenidae. Maps are aligned on two Sac II sites (positions 676 and 2356 bp), which are invariant throughout most of the Vertebrata (Carr *et al*, 1987).

* Arnason *et al* (1991), have since sequenced the entire mtDNA genome of the Fin whale, determining its length to be 16398 bp long.

The maps are orientated using a similar invariant Hpa I site which occurs at 5540 bp and therefore lies 3184 bp to the right of the second Sac II site.

10.5 INDIVIDUAL SITE ALIGNMENTS

Individual site alignments are illustrated in Figure 17, which provides an easier visual identification of alignments and hence relationships between individual taxa.

10.6 CLADISTIC AND DISTANCE MEASURE PHYLOGENIES

The fifteen enzymes used provide sixty-nine phylogenetically informative characters for cladistic analysis, i.e. those shared by at least two and by not more than n-2 (where n = sample size) taxa. The site positions giving rise to these characters and the informative character states are listed in Table III.

There are one hundred and twenty-six separate sites, of which ten are present throughout all the sampled taxa. The pairwise sequence divergence grid used for the construction of dendrograms is shown in Table IV.

10.7 CLADOGRAMS

Cladograms are shown in Figure 18, with (a) showing the single most parsimonious tree produced by the implicit enumeration options of Hennig 86 (J.S. Farris, version 1.5). Figure 18 (b) illustrates a Bootstrap consensus tree with 1000 replicates (Bootstrapped mixed parsimony algorithm version 3.1), which has an identical topology to the single most parsimonious tree.

10.8 DENDROGRAMS

Figure 18 (c) shows a dendrogram constructed using the Neighbor-Joining method (Saitou and Nei, 1987). The Fitsch-Margoliash (1967) method gives the same topology.

The cladograms and dendrograms have identical topologies.

(For Figure 16 please see overleaf.)

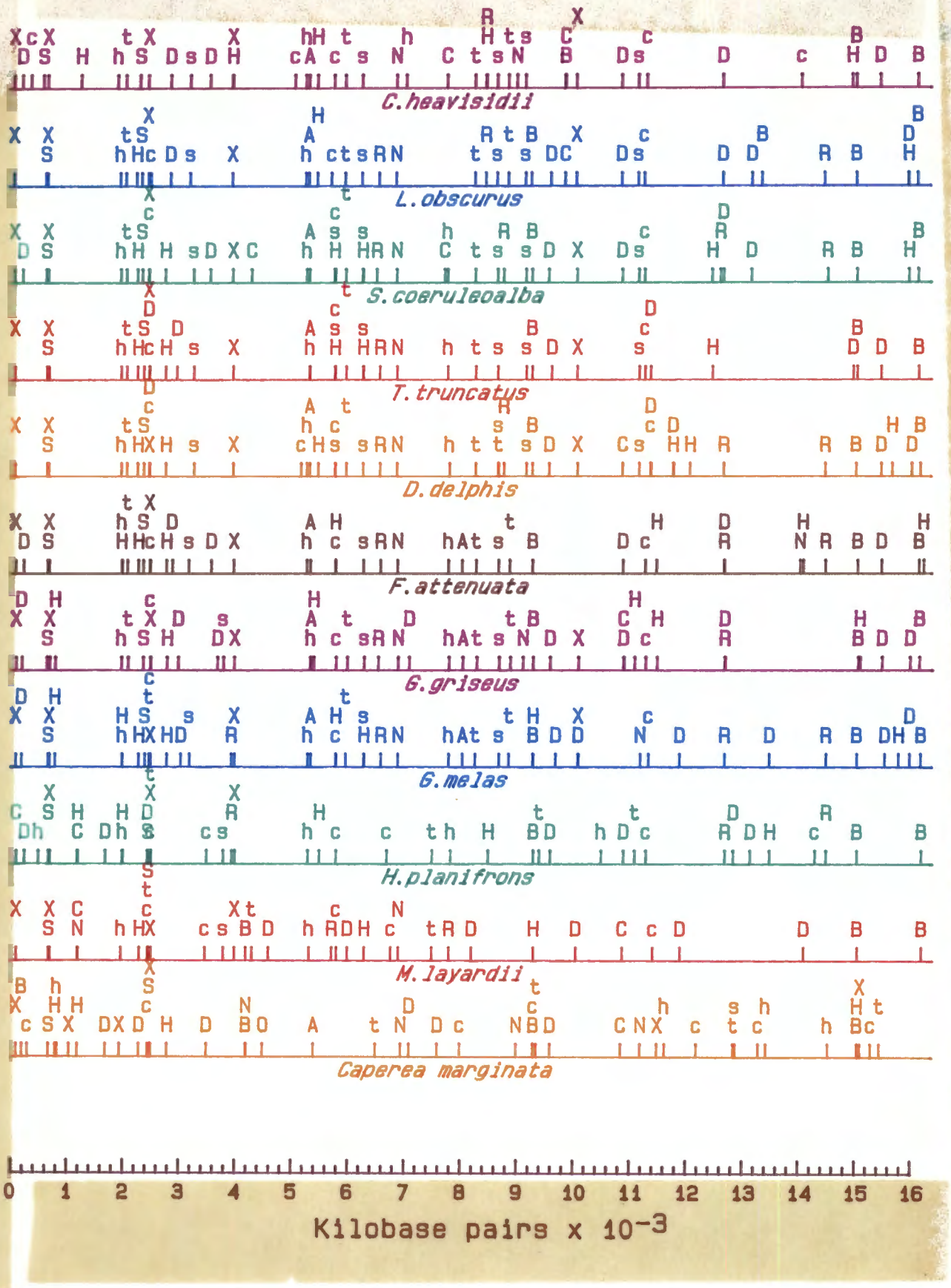


Figure 16 : Restriction endonuclease maps of mitochondrial DNA from eleven species of cetaceans are linearised and aligned on an invariant Sac II site. For enzymes used and codes refer Table II.

Figure 17 : Diagrammatic representation of the site alignments for each enzyme.

TABLE III

Informative site positions and character states

```
Site number                 1111111111222222222233333333334444444444555555555566666666666
                   123456789012345678901234567890123456789012345678901234567890123456789
Enzyme code        ssssssshAACCCCBBtttttttNNRRRRRRXccccccDDDDDDDDDDDDDDDDHHHHHHHHHHHHHHH
                         1      1          1    111    11        11111111          11111
                   33568917581790492257889913688240 35611 122337900112356 12225568912556
Position           288372283027982314952830295586513417342749061609497150820385835355070
                   500000005507080007909975058009005853113080100170000901000000000000550
                   700000085005000000460655500500180050520303060045000008200000000000000
                   ---------------------------------------------------------------------
Caperea marginata  000000001000011100000001110000000100000010001100000000110010000000100
H.planifrons       010000010010000101010010010001100101101110000101000100011001001000000
M.layardii         010000000010011001010000000000000010101000000010010000000100010100000
G.melas            100110011100000101101100111001110000101000100010010011101110110100010
G.griseus          010110011100010110101101001001010000101000101101001011100010000010100
F.attenuata        100110011100000110001100001001100000101001010001001010001110100010000
C.heavissidi       100111101001100010101101000100011010101001010001001010000001001000100
L.obscurus         100111101000100110101100001100110000100001000101001101000101000000001
S.coereoulba       101111111001000110101000001011110000101000010101001100000110110001001
T.truncatus        101111111000000110101000001000010000100010100100100010000110110001000
D.delphis          101111111000010110101000000101111001001001000010010001100011100000010
```

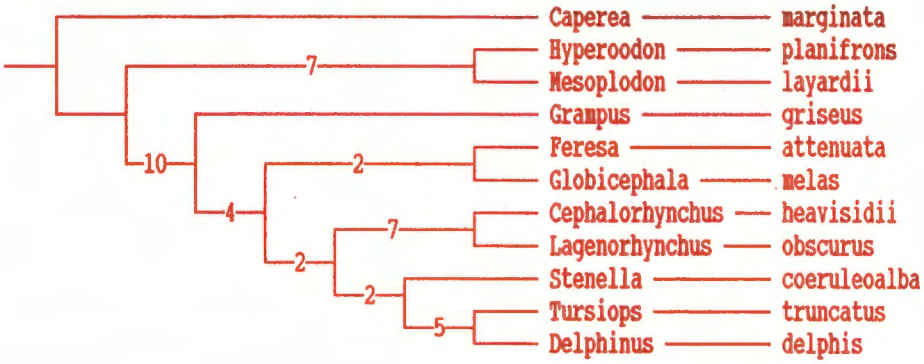
Site positions giving rise to the phylogenetically informative characters and the informative character states. The table is read column wise. For example, Informative Site No 1 is produced by Sac I (s), lies at position (base pairs) 3257 on the mtDNA genome and is present in the taxa G.melas through to D.delphis (O = not present).

TABLE IV

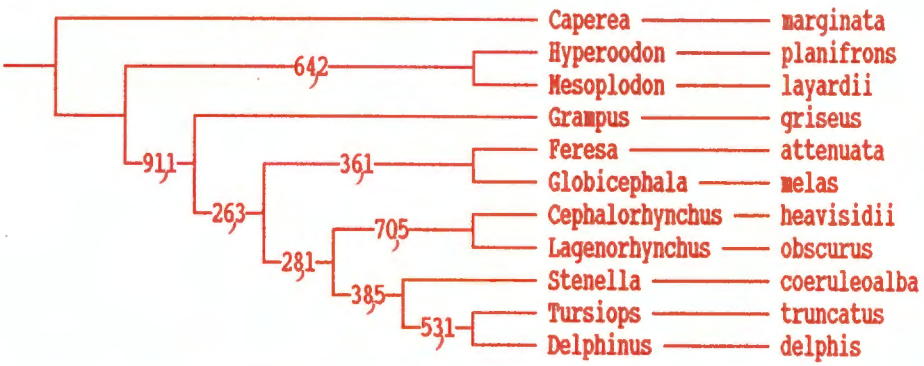
		1	2	3	4	5	6	7	8	9	10	11
1	Caperea marginata	45	.256	.225	.283	.256	.356	.270	.235	.262	.267	.264
2	H.planifrons	23.1	45	.506	.468	.477	.478	.413	.455	.395	.478	.426
3	M.layardii	25.4	11.4	38	.483	.395	.400	.353	.395	.430	.400	.414
4	G.melas	21.4	12.7	12.2	49	.739	.688	.521	.630	.711	.688	.694
5	F.attenuata	23.1	12.4	15.6	5.0	43	.756	.622	.674	.714	.756	.652
6	G.griseus	17.4	12.4	15.4	6.2	4.6	47	.660	.711	.659	.681	.667
7	C.heavissidi	22.2	14.9	17.6	10.9	7.9	6.9	47	.689	.614	.638	.583
8	L.obscurus	24.6	13.2	15.6	7.7	6.6	5.7	6.2	43	.714	.756	.696
9	T.truncatus	22.7	15.6	14.2	5.7	5.6	6.9	8.1	5.6	41	.818	.800
10	S.coereoulba	22.4	12.4	15.4	6.2	4.6	6.4	7.5	4.6	3.3	47	.750
11	D.delphis	22.6	14.4	14.8	6.1	7.1	6.7	9.0	6.0	3.7	4.8	49

Pairwise sequence divergence grid used for the construction of dendrograms. Figures below the diagonal are pairwise sequence divergence in percent. Figures on the diagonal are number of informative sites used per taxa. Figures above the diagonal are pairwise proportion of shared sites.

a) Maximum Parsimony - Length 150



b) Bootstrap Consensus tree



c) Neighbour-joining distance dendrogram

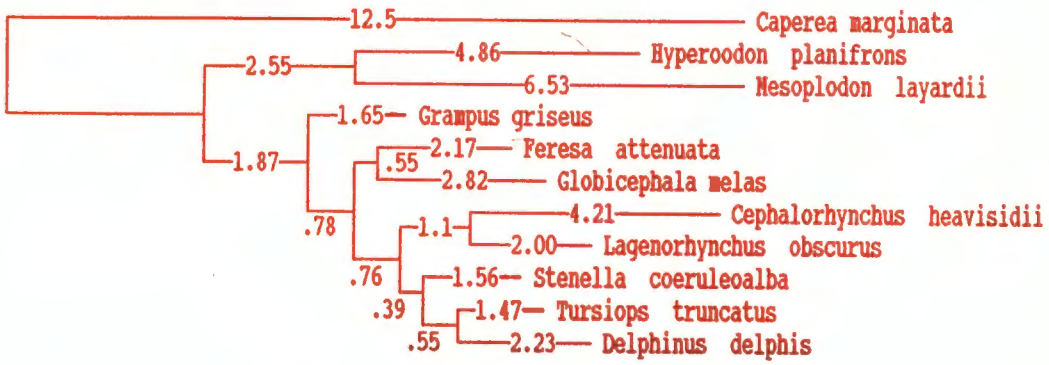


Figure 18 : Cladograms produced from the table of phylogenetically informative sites, with (a) the single most parsimonious tree produced by the implicit enumeration options of Hennig 86 (the figures indicate the number of symplesiomorphies that support the given cluster) and (b) Bootstrap analysis using the Bootstrapped mixed parsimony algorithm version 3.1 with 1000 replicates (the figures indicate in percentages the strength of the most frequent clusters obtained). (c) Distance dendrogram produced from the table of pairwise sequence divergences using the Neighbor-Joining method (the figures indicate the pairwise sequence divergence, in percentages, between taxa).

CHAPTER 11

DISCUSSION

11.1 FEATURES TO BE DISCUSSED

The phylogenetic reconstruction at the generic level of eight members of the Delphinidae using cladistic (maximum parsimony and bootstrap consensus tree methods) and distance (neighbor-joining [Saitou *et al*, 1987] and Fitch-Margoliash (1967) methods) approaches is undertaken. The results are compared to similar studies based on morphological and allozyme methodologies.

Using distance data an attempt is made to estimate the time of divergence at the (a) generic; (b) familial; and (c) suborder levels. The results are discussed with reference to palaeontological data, and with a study of the rate of neutral nucleotide substitution of cetacean nDNA and with the allozyme based study of Shimura *et al* (1987). Times of divergence are estimated using the mammalian mtDNA molecular clock (Brown *et al*, 1979), and are compared with a nucleotide base substitution rate for cetacean mtDNA, calibrated using the odontoceti-mysticeti palaeontologically estimated time of divergence.

11.2 STRUCTURE

In an attempt to maintain a concise level of discussion this thesis is so structured that lengthy evaluations of statistical methodologies, morphologically based phylogenies and palaeontological findings are wherever possible treated separately. Where it is felt necessary to substantiate the discussion, referrals to the relevant sections are made.

11.3 LIMITATIONS OF THE PRESENT STUDY

The limitations of this study are defined by the availability of biological material, which was primarily obtained through opportunistic strandings or accidental death in commercial trawl nets (refer Table 1). Nevertheless, taxa from eight of the seventeen genera which are variously grouped within three of the six subfamilies of Delphinidae were sampled. Of the taxa sampled, *Feresa attenuata, Delphinus delphis, Tursiops truncatus* and *Grampus griseus* are the only members of their genera. *Globicephala melas* is one of two species in this genus, whereas *Cephalorhynchus heavisidii, Stenella coeruleolba* and *Lagenorhynchus obscurus* have four, five and six members respectively within their genera (Perrin, 1989).

11.4 THE PHYLOGENETIC RECONSTRUCTION AT THE GENERIC LEVEL OF EIGHT MEMBERS OF THE DELPHINIDAE

11.4.1 Inferred Phlogenies

The constructed cladograms (i.e. the single most parsimonious tree and the bootstrap consensus tree using 1000 replicates) and dendograms (using the neighbor-joining and Fitch Margoliash methods) all have identical topologies.

The present study's inferred phylogeny supports a monophyly for *Lagenorhynchus, Tursiops, Delphinus, Cephalorhynchus* and *Stenella* (refer Figure 18). This is concordant with the morphologically based classifications of Kasuya (1973), Mead (1975), Fraser and Purves (1960) and Perrin (1989), who place these genera (amongst others) within the subfamily Delphininae (refer Chapter 3 for details).

The inferred between genera relationships within the subfamily Delphininae differ slightly from Perrin's phenogram (Figure 4). Although *Lagenorhynchus* is commonly agreed to be the most distantly related genus within the Delphininae (with reference to the sampled taxa), the present study finds *Tursiops* to be more closely related to *Delphinus*, whereas Perrin places *Stenella* in closer proximity to it.

In contrast to the cited morphological classifications which group *Cephalorhynchus* under a different subfamily from the Delphininae (i.e. either under Sotalinae or Cephalorhynchinae), this study's inferred phylogeny groups it with *Lagenorhynchus* under the Delphininae. This unexpected topology cannot be explained by the high frequency of autapomorphies in *C. heavisidii's* lineage (Figure 18(d)), as using Hennig's XX function (which is used to determine the 'strength' of a clustering) it was established that a substantial seven synapomorphies group the two together (cladistic analysis ignores autapomorphies). The bootstrapped mixed parsimony algorithm also supports a relatively strong pairing of *Lagenorhynchus* and *Cephalorhynchus* (705 of the 1000 replicates support this topology). It would be interesting to include *Phocoena* (family Phocoenidae) to see if this causes *Cephalorhynchus* and *Phocoena* to group independently of *Lagenorhynchus* (*Cephalorhynchus* has been noted by Mead (1975), Mitchell (1970) and Barnes (1978) to closely resemble *Phocoena* (porpoise). In an allozyme study by Shimura *et al* (1987) (Figure 5) which groups *Stenella* and *Tursiops* together, the dendogram indicates subfamily status for *Lagenorhynchus*. Of the taxa sampled in the present study, *Lagenorhynchus* is considered to be most closely related to *Cephalorhynchus*.

Grampus is the most basally placed taxon in the Delphinidae. If this represents the true phylogeny then it would imply that the Delphininae and Globicephalinae were a later radiation and a product of an earlier *Grampus* radiation.

It is not too difficult to reconcile this with the "common agreement" (Gaskin, 1982, p.179) that *Tursiops* is anatomically the most generalised of modern delphinids. A highly derived form can be young; there is not a one-to-one correspondence between morphologic and genetic difference. A generalized condition can be a plesiomorphy. *Grampus* can be highly derived and still be descended from a form (presumably more generalized) closer to the outgroup genetically than *Tursiops*. The results suggest that *Grampus* might be most properly placed in its own subfamily.

The latest Miocene deposits of California have produced fossil remains that resemble modern *Stenella* and *Tursiops*, and a probable globicephalid from the late Pliocene has been excavated (Barnes, 1976). The morphologically based classifications place *Grampus* either under the Globicephalinae (Kasuya, 1973) or under the Delphininae (Mead, 1975, Fraser and Purves , 1960 and Perrin, 1989).

Feresa and *Globicephala* are inferred to assume a close genetic relationship. This is consistent with all four of the cited morphological classifications (i.e. Kasuya and Perrin group them under the Globicephalinae, whereas Mead and Fraser and Purves group them under the Orcinae (refer Chapter 3).

11.4.2 Critical Analysis of the use of Cladistic and Distance Methods in the Interpretation of Molecular Data

Distance methods are criticised on the basis that information is lost during the transformation of the raw data into distance matrices, whereas cladistics is based on a direct raw-data analysis of transformative character states.

However, distance measures use more of the characters than do cladistic methods when considering a molecular data base (Li *et al*, 1991), since they utilise the autapomorphic sites. For example, restriction endonuclease maps constructed from members of the family Delphinidae in the present study have 126 individual sites used by distance measures, whereas only 69 informative sites for cladistic analysis.

For morphological data cladistic methods are clearly superior, since the rate of accumulation of morphological characters is poorly correlated with time. For molecular data the situation is somewhat different. Relative rate tests (Sarich and Wilson, 1973) and calibration of the molecular clock against the fossil record (Brown *et al*, 1979) generally uphold the constancy of the rate of accumulation of neutral mutations in DNA, especially in mtDNA, at least within closely related groups such as families, such that the principal error in relative rate measurements is due to the stochastic nature of mutational events.

On the other hand, the frequency of parallel and back mutations in DNA, especially where there is a high transition bias (which is > 90% in mtDNA), make for a high frequency of homoplasis. Thus for molecular data, distance methods are stronger and cladistic methods relatively weaker, as compared with morphological data.

In support of distance measures based on sequence divergence for inferring phylogenies, Takahata *et al* (1991) have demonstrated (by using formulas for the minimum and maximum values of the sampling variance) the (statistically) satisfactorily accurate estimates of the sampling variance of nucleotide diversity as used by distance methods.

Jin *et al* (1991) compared the relative efficiencies of the Maximum Parsimony (MP) and Neighbor-Joining (NJ) methods in obtaining the correct topology for restriction site data. Computer simulations assumed the rate of nucleotide evolution to follow a given model tree. Restriction site data (obtained from the recognition sequences of 20 six-base restriction enzymes) was used to reconstruct a phylogenetic tree which was compared to the model tree. The results showed that the probability of obtaining the correct tree is higher for the NJ method than for the MP method if the rate of nucleotide substitution is constant. If the average topological deviation from the model tree is used as the criterion, both methods are almost equally efficient. However, when the rate of nucleotide substitution varies with evolutionary lineage, the NJ method proved superior to the MP method in estimating the true topology.

Cladistic methods, if using a large number of informative sites, are still probably the best way to estimate the true topology, but distance methods have their main value in estimating the timing of interesting biogeographic or evolutionary events. Hence each approach has its place in the interpretation of comparative molecular data.

11.4.3 Estimated Times of Divergence at the Suborder, Familial and Generic Levels of Cetacea using Distance Data

11.4.3.1 Estimated times of divergence based on the mammalian mtDNA molecular clock

The most significant feature to emerge from this study is the evidence from the distance data of the very recent radiation at the generic, familial and suborder levels of Cetacea when compared to palaeontological estimates.

Assuming for the present that the calibration of the mammalian mtDNA molecular clock originally established by Brown *et al* (1979) is valid for cetaceans, then :-

Generic level

The major radiation at the generic level within the Delphinidae took place about two million years ago (start of the Pleistocene). The two ziphiid genera, *Hyperoodon* and *Mesoplodon*, diverged about 5 - 6 Myr ago (early Pliocene).

Familial level

At the familial level the Ziphiidae and the Delphinidae are calculated to have diverged about 7 - 8 Myr ago.

Suborder level

At the suborder level the odontoceti and mysticeti diverged as recently as 11 - 12 Myr ago (middle Miocene).

11.4.3.2 Comparative analysis of estimated times of divergence

A cursory comparison of this study's estimated times of divergence with Barnes' palaeontologically-based estimates at the family (Delphinidae and Ziphiidae split 25 - 30 Myr ago) and suborder (odontoceti and mysticeti split 35 - 40 Myr ago) levels of Cetacea (Figure 3) leads one to make the following deductions :-

Assuming that the fossil-based estimates are correct, then the cetacean mtDNA's base substitution rate is significantly slower than the calculated 1% substitution rate of Brown *et al*'s (1979) mammalian mtDNA molecular clock. If the calibration of the mammalian mtDNA clock is valid for cetaceans then the phylogeny of cetacean families needs to be reappraised in that radiation of modern whales occurred much more recently than is currently estimated.

Forms similar to modern cetaceans would then be examples of parallel evolution, rather than representative of modern cetaceans' plesiomorphic states.

A confounding factor which prevents an unambiguous evaluation of the above dichotomy is the sparsity of fossil evidence (refer Chapter 2), which -

(a) makes it difficult to construct a definitive palaeontologically based phylogeny; and

(b) makes the calibration of the molecular clock specific to cetaceans quite hazardous (refer Chapter 8).

11.4.3.3 Suborder level - estimated times of divergence of the odontoceti and mysticeti

Barnes (1984), Gaskin (1982) and Whitmore and Sanders (1976), contrary to Yablokov (1972) (Refer Chapter 3), convincingly support the monophyletic status of the two extant suborders of Cetacea.

Gaskin (1982) considers the mysticeti to be a monophyletic group on the grounds that it would be difficult to conceive of the parallel evolution of such a unique feeding strategy such as the mysticeti have developed.

If the above authors are correct, then the earliest fossil evidence of (modern) odontocetes and mysticetes (which shared the Archaeoceti as a common ancestor) must be indicative of the time of divergence, or earlier, of the two extant suborders.

Similarly, if the mysticeti (as Gaskin supposes) do not have any parallel lineages then the earliest mysticeti fossil records should indicate the time of divergence, or earlier, of the mysticeti from the odontoceti. The earliest mysticeti taxon yet discovered is the now extinct *Cetotheriidae*, and is dated at approximately 35 Myr (mid-Oligocene) (Barnes *et al*, 1985).

These arguments convincingly estimate the time of divergence between the two extant suborders at around 35 - 40 Myr ago, which is quite incompatible with the present study's molecular estimate (using Brown *et al*'s molecular clock), which positions the divergence at approximately 12 Myr ago.

However, according to Barne *et al*'s (1985) phylogeny of cetacean families, there are no fossil records which directly link the Archaeoceti with the modern cetacean suborders, nor is there any direct fossil evidence of the common point of divergence of the mysticeti from the odontoceti. That the mysticeti evolved from some primitive type of odontoceti is probable (refer Chapter 2), but as mentioned the only feasible method of estimating the time of divergence of the two suborders (assuming the monophyletic status of the mysticeti) is to use the earliest known mysticeti fossil records as the latest estimated time of divergence.

The question remains as to whether baleen whales really are monophyletic. Assuming the validity of Brown *et al*'s molecular clock for cetaceans and the authenticity of palaeontological deductions, then the only resolution to the incompatible molecular and fossil time of divergence estimates would be for the mysticeti to be paraphyletic.

Then it could be postulated that *Caperea marginata* (family Neobalaenidae), for which virtually no fossil evidence exists, evolved more recently.

It is interesting to note that the pairwise sequence divergence percentages between the baleen and toothed whales are very similar, indicating that the baleen whale (*C.marginata*) and the sampled odontocetes (from families Delphinidae and Ziphiidae) shared a common ancestor.

11.4.3.4 Neutral substitution rate of nDNA

Analysis of the neutral nucleotide substitution rate for cetacean nDNA by Schlötterer *et al* (1991) supports the present study's findings of the either very slow rate of substitution in cetacean nDNA and mtDNA or that the fossil record needs to be re-appraised. Schlötterer *et al* (1991) calculate the rate of neutral nucleotide substitution for cetacean nDNA at 0,09% Myr-1, compared to the average typical divergence rate for neutral nucleotide positions of 0,5% Myr-1 (Wilson *et al*, 1987).

As in the present study, Schlötterer *et al* used the mysticeti/odontoceti split to calculate the base substitution rate. Their calculated figure represents the slowest substitution rate found so far for any taxon.

Secondly, their data supports the palaeontologically-based reconstruction (Barnes, 1985) of the evolution of the major odontoceti families and the mysticetes at around the same time (as the average among the odontocete families is similar to that between the odontocetes and mysticetes) (refer Section 3.6.2).

11.4.3.5 Family level - estimated times of divergence of the Delphinidae and Ziphiidae

Fordyce (1980) dates the earliest ziphiid fossils at 25 Myr ago (start of the Miocene), whereas Barnes (1985) makes an earlier estimate of 18 - 20 Myr ago (early to mid-Miocene). Barnes (1984) and Gaskin (1982) date the earliest Delphinidae fossil remains at 10 Myr, whereas Fordyce (1980) records the earliest known Delphinidae fossils at the mid-Miocene. A confounding factor in the estimation of the age of the extant family Delphinidae is that other delphinid-like families existed in the early Miocene, which may or may not be representative of parallel lineages (refer Chapter 3).

Barnes (1984) and Gaskin (1982) both deduce from the available fossil evidence that the extant family Delphinidae evolved from the extinct Kentriodontidae. The Ziphiidae and Kentriodontidae are thought to have shared a common ancestor from the mid-Oligocene (30 - 35 Myr) (Barnes, 1984).

The molecular estimated time of divergence between the Ziphiidae and Delphinidae, using the mtDNA molecular clock (Brown *et al*, 1979), is 7 - 8 Myr ago (late Miocene) !

Allozyme studies by Shimura *et al* (1987) using the species *Berardius bairdii* (Baird's beaked whale) estimate the time of divergence between Ziphiidae and Delphinidae to be either 3,5 - 5,5 Myr or 13,3 - 20 Myr ago, depending on which set of equations are used, resulting in an extremely broad estimate (refer Section 3.6.1).

Mead (1975) has stated that *Mesoplodon* and *Hyperoodon* may have derived from the advanced ziphiids *Belemnoziphus* or *Proroziphius* of the late Miocene (refer Chapter 2). To be compatible with the molecular datings this would require the Ziphiidae to be paraphyletic. Further molecular analysis on the ziphiid species should readily resolve this point.

The apparent recent divergence of the Delphinidae from the Ziphiidae is also more compatible with Heyning's (1989) contention that the Ziphiidae and Physeteroidea are not a monophyletic group within the super family Physeteridae, but that the Physeteridae diverged earlier from the lineage which led to the common ancestor of the Ziphiidae and the Delphinidae (refer Section 2.6.2).

11.4.3.6 Generic level - estimated times of divergence of genera within the family Delphinidae

Karyotype morphology of modern Delphinidae is quite uniform. The C-heterochromatin content varies only between 10 to 15% and the diploid number is 2n = 44 (Walen *et al*, 1965). Palaeontological evidence indicates that the Pliocene (2 - 5 Myr) was the period in which the modern delphinids became most abundant (Gaskin, 1982). However, the modern Delphinidae lineage probably extends back to the late Miocene (Barnes, 1984). Fossil remains resembling *Stenella* and *Tursiops* have been recovered from the latest Miocene deposits (Barnes, 1976 and Gaskin, 1982). Kellog (1931) has described the evolution of *Tursiops* from the early Pliocene. A probable *globicephalid* has been recovered from the late Pliocene (Barnes, 1976).

The estimated 2 Myr time of generic divergence of the family Delphinidae of the present study is not too inconsistent with palaeontological findings depending on whether the late Miocene delphinid fossil remains represent the plesiomorphic state for the Delphinidae or whether they are an example of an earlier parallel evolutionary radiation.

Cephalorhynchus has been noted by Mead (1975), Mitchell (1970) and Barnes (1978) to closely resemble *Phocoena* (porpoise). Mead based his conclusions on facial anatomy similarities and is of the opinion that this possible convergence of *Cephalorhynchus* to the *Phocoena* pattern justifies a possible subfamily status for the *Cephalorhynchus* spp. Using both distance and cladistic methods the present study reveals *Cephalorhynchus* to be as shallowly rooted (i.e. as recently evolved) as any of the other Delphinidae, a topology which would not support its proposed subfamily status. The Neighbor-joining distance dendogram (Figure 18(c)) shows an abnormally high degree of autapomorphies (4,32% compared to approximately 2% pairwise sequence divergence between the other dolphins), a figure which reflects *Cephalorhynchus*' uniqueness to the other sampled members of Delphinidae. However, cladistic analysis (which does not use autapomorphic characters), supports the *Lagenorhynchus / Cephalorhynchus* grouping by a substantial seven steps (analysed using Hennig's XX function).

The inclusion of a *Phocoena* in future phylogenetic studies should readily resolve the postulated convergence of *Cephalorhynchus* towards *Phocoena* characteristics.

11.4.3.7 Comparison with Allozyme Studies

The only comprehensive allozyme study to date on the Delphinidae is by Shimura *et al* (1987). An accurate conversion of genetic difference into time is confounded by the two quite divergent estimates arrived at by using different statistical approaches.

Using the equations of Nei (1975), Shimura *et al* arrive at extremely recent divergence time estimates at the generic level of Delphinidae (approximately 1 Myr ago) (Figure 5). Using a second group of calculations based on albumin immunological distance (AID), they estimate the time of generic divergence at 7,6 Myr ago. Of these two divergent estimates the one based on Nei's equations agree well with the present study's findings.

11.4.3.8 Calculation of Cetacean mtDNA Base Substitution Rate

Assuming that the palaeontologically estimated time of divergence of the mysticeti and odontoceti is correct, then using the pairwise sequence divergence percentages from the distance data, the base substitution rate of cetacean mtDNA can be calculated as follows :-

Average sequence divergence rate between the odontoceti and mysticeti is 22%
Estimated time of divergence is 35 - 40 Myr
Therefore the base substitution rate is 11% over 35 - 40 Myr, or 0,3% per Myr.

This figure is significantly lower than the estimated 1% nucleotide base substitution rate for mammalian mtDNA as calculated by Brown *et al* (1979), and is concordant with the low neutral substitution rate of nDNA as calculated by Schlötterer *et al* (1991).

11.4.3.9 Possible explanations for the low substitution rate of cetacean nDNA and mtDNA

Generation times

Differences in nucleotide substitution rate have been correlated with the varying generation times among groups of taxa (refer Chapter 8). However, the generation times of cetaceans are reasonably similar to those of primates (Kasuya, 1984), and as such cannot account for the significantly slower base substitution rate of cetaceans.

Non-mutagenic environment

Schlötterer *et al* (1991) propose that the slow substitution rate of cetaceans could be attributed to the less mutagenic ocean environment.

For example, the influence of cosmic rays is much less for cetaceans than for terrestrial mammals, as their effect is already reduced by 70 % at 10 m below sea level. Neither are cetaceans subjected to the largest source of irradiation, the soil.

Finally it can be postulated that marine environs are possibly more stable than terrestrial ones, thus lessening the necessity of continual adaption. This fact possibly affects mtDNA sequence divergence estimates, as both coding and non-coding regions are sampled using this technique.

11.4.3.10 Re-appraisal of fossil evidence

The alternative explanation to the incompatible molecular and palaeontological time of divergence estimates is that the earlier, similar forms of modern cetaceans do not represent their plesiomorphic states, but rather are examples of earlier parallel evolution and that therefore the extant cetacean families evolved much more recently than is currently supposed. The poor fossil record of cetaceans makes this a feasible hypothesis, as fossils older than the early Miocene (20 - 25 Myr ago) cannot unequivocally be associated with the extant families' lineages (refer Chapter 2).

11.5 CONCLUSION

Restriction endonuclease site mapping (RSM) has proved to be a viable molecular technique for inferring phylogenies, as using this technique sampled cetaceans were successfully grouped under suborder, family and subfamily levels.

The generic relationships within the three subfamilies of Delphinidae are concordant with morphologically based classifications. The two main differences, viz. *Grampus*' most basally rooted position and *Cephalorhynchus*' grouping with the Delphininae are of taxa whose groupings are unresolved in morphologically based classifications. Although it remains a possibility that the inferred phylogeny may change with the addition of other taxa, it seems reasonable to assume its good approximation to the true one as both cladistic and distance measures produced identical topologies.

The low pairwise sequence divergence figures obtained, translated into real time using Brown *et al*'s (1979) molecular clock, has resulted in very recent divergence dates at the generic, family and suborder levels when compared to palaeontologically based estimates. It remains an unresolved issue as to whether the base substitution rate of cetacean DNA is substantially slower than that of terrestrial mammals or whether the fossil evidence needs to be re-interpreted. The possible inaccuracy of the molecular clock (primarily due to the stochastic nature of mutational events and difficulties in calibration), and the incomplete fossil record for cetaceans precludes a definitive conclusion from being made.

That the base substitution rates of nDNA and mtDNA of cetaceans have been measured to be significantly slower than that of terrestrial mammals, possibly supports the need to re-appraise the fossil evidence as nDNA and mtDNA evolve independently of each other. However, firstly the base substitution rate for both nDNA and mtDNA was calculated using the mysticeti/odontoceti split, the time of which is palaeontologically uncertain. Secondly the monophyletic status of the two extant suborders is not definite, thus making the calculation of base substitution rates risky.

APPENDIX I

MATERIALS

---

1. ISOLATION AND PURIFICATION OF MITOCHONDRIAL DNA

1.1 Buffers

(a) Extraction Buffer (1 M)

To make 1000 ml: 100 mM Tris.Cl (pH 8)
150 mM NACl
20 mM EDTA (pH 8)
10% w.v sucrose

Add distilled H_2O to 1000 ml. Autoclave.

(b) Tris EDTA (TE) Buffer (1 M, pH 8)

To make 1000 ml: 10 mM Tris.Cl (pH 8)
1 mM EDTA (ph 8)

Add distilled H_2O to 1000 ml. Autoclave.

(c) Saline Tris EDTA (STE) Buffer (1 M, pH 8)

To make 1000 ml: 10 mM Tris.Cl (pH 8)
1 mM EDTA (pH 8)

Add distilled H_2O to 1000 ml. Autoclave.

(d) Tris.HCl Buffer (1 M, pH 8 or pH 7,5)

To make 100 ml: Dissolve 121.1 g Tris base in 800 ml distilled H_2O. Adjust pH to 8 with conc. HCl. Autoclave. If pH 7.5 required, add another 20 ml conc. HCl before making up to 1000 ml.

1.2 Miscellaneous Solutions

(a) Ethylene Diamino Tetra-Acetic Acid (EDTA) (0,5 m, pH 8)

To make 500 ml: Dissolve 93 g EDTA in 400 ml distilled H_2O. Adjust pH to 8 with 10 g NaOH and 5 M NaOH. Add distilled H_2O to 500 ml. Autoclave.

(b) Sodium Chloride (NaCl) (5 M Stock Soln.)

To make 500 ml: Dissolve 146 g NaCl in 400 ml distilled H_2O. Add H_2O to 500 ml.

(c) Sodium Dodecyl Sulphate (SDS) (10% Stock Soln.)

To make 100 ml Add 10 g SDS to 90 ml distilled H_2O. Warm to dissolve, and make up to 100 ml with H_2O.

(d) Sodium Hydroxide (NaOH) (5 M Stock Soln.)

To make 500 ml: Dissolve 100 g pellets NaOH in 400 ml distilled H_2O. Add dH_2O to 500 ml.

(e) Ethidium Bromide (EtBr) (10 mg/ml Stock Soln.)

To make 100 ml: Add 1 g EtBr to 100 ml distilled H_2O. Stir to dissolve. Wrap container in aluminium foil and store at 4 °C.

2. RESTRICTION ENZYME DIGESTION

2.1 Buffers

(a) KGB Buffer (2 x Stock Soln.)

To make 1000 ml:
200 mM Potassium Glutamate
50 mM Tris Acetate (pH 7,6)
20 mM Magnesium Acetate
100 g/ml Bovine Serum Albumin
1 mM 2- Mercapto-Ethanol

To make 20 ml:
0,741 g mM Potassimu Glutamate
1 ml 1M Tris Acetate (pH 7,6)
0,090 g Magnesium Acetate
1 ml 2 mg/ml BSAI Soln.
400 l 50 mM 2- M EtOH

Add sterile distilled H_2O to 20 ml. Filter sterilize. If a 1 x conc. is required, dilute 1/2 before use.

2.2 Miscellaneous Solutions

(a) Tris Acetate (1 M, pH 7.6)

To make 1000 ml: Dissolve 121,1 g Tris base in 800 ml distilled H_2O; adjust pH to 7.6 with Acetic Acid. Make volume up to 1000 ml. Autoclave.

(b) Restriction Enzymes

Enzymes were diluted in 1 x KGB Buffer from laboratory stocks, to a working concentration of 2 units/ l per digest.

3. END-LABELLING REACTIONS

3.1 Reagents Used

(a) Deoxynucleotides

Each of the three Deoxynucleotides, dATP, dTTP and dGTP, were diluted in sterile distilled H_2O from laboratory stock solutions of concentration 20 mM, to a final concentration of 2 mM per end-labelling reaction.

(b) ^{32}P DeoxyCytidine Phosphate

^{32}P dCTP was diluted in sterile distilled H_2O from a laboratory stock solution of 10 Ci/ l, to a working concentration of 1 Ci/ l per end-labelling reaction.

(c) Klenow Polymerase

Klenow polymerase was diluted in sterile distilled H_2O from a laboratory stock solution of 6 units/ l, to a working concentration of 1 unit/ l, per end-labelling reaction.

4. GEL ELECTROPHORESIS

4.1 Buffers

(a) Tris Acetate EDTA (TAE) Buffer (50 x Stock Soln.)

To make 1000 ml: 242.0 g Tris base
57.1 ml Glacial Acetic Acid
100.0 ml 0,5 M EDTA 9pH 8)
0.05% v/v Sodium Pyrophosphate

Add distilled H_2O to 1000 ml. Autoclave. Store at room temperature. Dilute 1/50 before use.

(b) Gel Loading Buffer (6 x)

To make 30 ml: 0,25% w/v Bromophenol Blue
40,00% w/v Sucrose
20 mM EDTA (pH 8)

Add distilled H_2O to 30 ml. Store at 4 ℃.

4.2 Miscellaneous Solutions

(a) Sodium Pyrophosphate (NaPP) (10% Stock Soln.)

To make 1000 ml: Dissolve 100g Sodium Pyrophosphate in 1000 ml distilled H_2O.

(b) Preparation of Large Agarose Gels for Electrophoresis

For a typical 1.2% gel, dissolve 1,8 g dry agarose powder in 150 ml 1 x TAE Buffer, by heating in microwave oven. When cooled to about 50 °C, pour onto gel apparatus and allow to set.

APPENDIX II

Restriction fragment sizes (kb) of Lambda DNA digested with Hind III :

23.130

9.416

6.682

4.361

2.322

2.027

0.564

0.125

APPENDIX III

Instructions for RESOLVE

Version 2.0. Copyright reserved 1990, E.H. Harley, Dept. of Chemical Pathology, University of Cape Town, Observatory 7925, Cape, South Africa.

RESOLVE is a program to help in the mapping of restriction endonuclease sites in circular DNA molecules, as well as to store, manipulate, and edit the maps produced, and to analyse the maps either for phylogenetically informative sites or for preparation of pairwise sequence divergence matrices. The program is designed to be user-friendly with little requirement for constant reference to the instructions.

The program is written in True BASIC, a compiled and structured derivation of the BASIC programming language. It runs on any IBM-compatible PC but performs best on an AT with a colour monitor.

The program has three main functions:

1. To construct restriction maps of DNA molecules from double-digestion data. This is performed in a two-step process to maximise the validity of the final map and to resolve the ambiguities and redundancies which result from a single set of single double-digestion data using two restriction enzymes.
2. To catalogue, edit, manipulate and manage sets of mapped data from a number of different DNAs.
3. To perform comparisons of the maps where these are from related DNA species (e.g. sets of mitochondrial DNA from related taxa) and to construct output files of either phylogenetically informative sites or sequence divergence matrices, appropriately formatted for a number of currently popular phylogenetic analysis programs (e.g. PAUP, HENNIG 86, Neighbour-joining, PHYLIP etc).

The Opening (Main) Menu:

This is displayed as follows:

MAIN MENU

1. Display management file status
2. Display temporary solutions
3. Display final maps
4. Edit DNA, or restriction enzyme files
5. Edit temporary solution files
6. Edit final map files
7. Map new enzymes to temporary files
8. Three enzyme consensus analysis (temp. to final maps)
9. Analyses of final maps
10. Toggle display/print modes
11. Exit program

Option 1. Displays the management file status which lists the DNA molecules being mapped, details of 2-enzyme double-digest solutions in the temporary file, and enzymes which have been mapped fully in the final file. The 2-enzyme double digest solutions are given as a concatenated sequence of three letter sets, where the letters in the first two positions refer to the single letter codes you (or I) have chosen for each restriction enzyme : E for Eco R1, B for Bam H1 etc (note that upper and lower case letters are distinguishable), and the figure in the third position gives the number of map solutions found

Instructions for RESOLVE

Version 2.0. Copyright reserved 1990, E.H. Harley, Dept. of Chemical Pathology, University of Cape Town, Observatory 7925, Cape, South Africa.

RESOLVE is a program to help in the mapping of restriction endonuclease sites in circular DNA molecules, as well as to store, manipulate, and edit the maps produced, and to analyse the maps either for phylogenetically informative sites or for preparation of pairwise sequence divergence matrices. The program is designed to be user-friendly with little requirement for constant reference to the instructions.

The program is written in True BASIC, a compiled and structured derivation of the BASIC programming language. It runs on any IBM-compatible PC but performs best on an AT with a colour monitor.

The program has three main functions:

1. To construct restriction maps of DNA molecules from double-digestion data. This is performed in a two-step process to maximise the validity of the final map and to resolve the ambiguities and redundancies which result from a single set of single double-digestion data using two restriction enzymes.
2. To catalogue, edit, manipulate and manage sets of mapped data from a number of different DNAs.
3. To perform comparisons of the maps where these are from related DNA species (e.g. sets of mitochondrial DNA from related taxa) and to construct output files of either phylogenetically informative sites or sequence divergence matrices, appropriately formatted for a number of currently popular phylogenetic analysis programs (e.g. PAUP, HENNIG 86, Neighbour-joining, PHYLIP etc).

The Opening (Main) Menu:

This is displayed as follows:

MAIN MENU

1. Display management file status
2. Display temporary solutions
3. Display final maps
4. Edit DNA, or restriction enzyme files
5. Edit temporary solution files
6. Edit final map files
7. Map new enzymes to temporary files
8. Three enzyme consensus analysis (temp. to final maps)
9. Analyses of final maps
10. Toggle display/print modes
11. Exit program

Option 1. Displays the management file status which lists the DNA molecules being mapped, details of 2-enzyme double-digest solutions in the temporary file, and enzymes which have been mapped fully in the final file. The 2-enzyme double digest solutions are given as a concatenated sequence of three letter sets, where the letters in the first two positions refer to the single letter codes you (or I) have chosen for each restriction enzyme : E for Eco R1, B for Bam H1 etc (note that upper and lower case letters are distinguishable), and the figure in the third position gives the number of map solutions found

for that particular enzyme pair. The sequence of letters (up to 20) under "R.E.S. in final file" are the individual single letter codes for enzymes which have been mapped rigorously by the three enzyme consensus analysis (option 8) or by editing into the final map files (option 6). When you first run the program after receiving the diskette you will see a set of DNA files with both partial and final map results on file. These are to provide a data set for practicing the various features of the program. It might be well to copy them to a "practice" directory. These files can be easily edited out (option 4) at any stage.

Option 2. Calls up the management file again and requests entry of a number corresponding to the DNA of interest. On entry of the latter a table of restriction enzymes currently in use is displayed and you are requested to choose a pair of numbers (enter each number, separated by a comma) corresponding to an enzyme pair with solution(s) in the temporary file. The map(s) will then be displayed, linearised at an arbitrary position (note that if sites are too close together a single letter code may be overwritten).

Option 3. Calls up the management file and requests a number corresponding to the DNA of interest. On entry of the latter the definitive map will be displayed in the upper half of the screen, linearised either at an arbitrary site or at a site chosen by editing (see option 6). Note that, as above, if sites are too close together then single letter codes will be overwritten. Below this graphical display will be listed the enzyme single letter codes and their mapped positions in the DNA sequence. Fragment sizes for any enzyme can be listed if required, or other maps listed under the first one for comparison.

Option 4. This calls up a secondary menu for editing of DNA or restriction enzyme files:

Edit menu for DNA or restriction enzyme files

1. Enter a new DNA
2. Delete a DNA file
3. Correct a DNA file name
4. Alter length of the DNA
5. Enter a new restriction enzyme
6. Delete a restriction enzyme
7. View restriction enzyme file
8. Escape to main menu

The options under this list enable entering, renaming, or deletion of DNA names and associated data files, or restriction enzymes. In the latter case you can construct a palette of enzymes and single-letter code names according to your requirements. The program comes with a palette of commonly used enzymes already on file, as well as a set of dummy DNA files to allow for practice with the various features of the program.

Option 5. This takes a rather low level look at the file containing the temporary double digest solutions. Each line in the list displayed shows a number giving the position in the file, the DNA name, the enzyme pair, and a string showing the order of enzyme positions in the temporary map. Sets of solutions of a particular enzyme pair will not necessarily be in adjacent positions. By entering either the letter 'L', 'D', or 'X' you can continue Listing

the data file, Delete one of the partial solutions, or eXit. This option is mostly used when temporary solutions have not been deleted after a three enzyme consensus analysis (you had been given the choice), but when you eventually have to clear these results to make way for new results. The size of the temporary data file is deliberately set to allow room for only a limited set of results - 12 in total - to discourage hoarding and encourage a parsimonious approach to mapping management.

Option 6. This calls up a secondary menu with various features for editing the final map data files :

Edit menu for map files

1. Delete whole map
2. Delete individual enzymes
3. Enter data for new enzyme(s)
4. Correct data values
5. Align map on an enzyme site
6. Reverse map orientation
7. Escape to main menu

Features 1 and 2 enable the whole map, or individual enzymes and their site positions, to be deleted.
Feature 3 enables you to edit in results from an enzyme or enzymes, which you have solved by hand or obtained by some other means. The results are integrated into the current final map. This is also the only means to document in the final file an enzyme with no cutting sites in the DNA, which is necessary if comparisons of related sequences are to be performed for phylogenetic purposes (option 9).
Feature 4 enable major or minor corrections to be made to site positions, and the order of sites will be adjusted if necessary.
Feature 5 realigns the whole map on a specified site position. This feature will always be necessary at some stage as the map is being built up, if comparisons with related maps are to be made, since the first map produced by option 8 will have an arbitrary starting point and further additions build on this. The usual approach is to identify invariant site positions (such as the Sac II sites in vertebrate mitochondrial DNA) and align maps on these.
Feature 6 enables the map site positions and order to be reversed and is useful in conjunction with site alignment (feature 5).

Option 7. This is the start of the whole mapping procedure using double digestion data. After selecting the DNA and the restriction enzyme pairs from the tables displayed, you enter the number of fragments given by each enzyme separately, followed by a likely value for the error (%) which is reasonable to expect from your measurements. This may require some trial and error : too low a value and no solutions are likely to be found, too high, and too many solutions, many clearly inappropriate, will be found. Since it is easy to try again later at a different error value without the tedium of re-entering all the data, several different values can be tried sequentially. Note that it is not always appropriate to use an error value giving only a single solution, since it is quite possible for another solution found at a slightly higher error value to be the correct one.

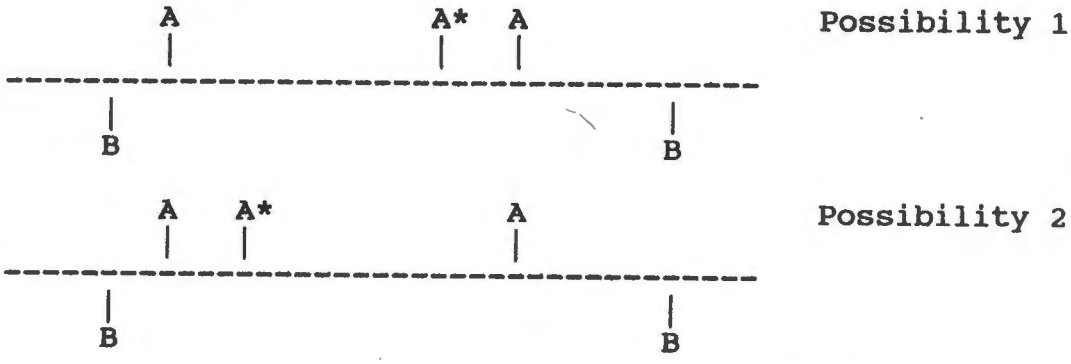
After choosing an error value the fragment sizes for the single and for the double digest results are entered. On completion of entry the values (sorted in descending order) and totals are displayed, and any

corrections can be made at this stage. An opportunity is also given to normalize the totals, and hence all fragment sizes, to a particular value. This is useful if one of the three sets of values is clearly out of line with the rest (e.g. where alignment with a marker lane on the gel was not optimal) and experience has shown that better results and tighter maps are often obtained when this option is used.

The program then searches for all ways in which some of the double-digest fragments (AB fragments) can fit single digest fragments (A or B fragments) within the error limits allowed. These matches rush past you on the screen and if you want a little more insight into the algorithm you can PAUSE to observe what is going on. Sets of partial solutions are gathered whereby all double digest fragments are matched singly or in combination (up to 6 double digest fragments per single fragment) with each of the two sets of single digest fragments, with no duplications or omissions. Each member of the set of 'A' partial solutions consists of sets of 'AB' fragments and is matched in turn with each member of the 'B' (second enzyme) sets. If a path is then found by which one can step from all A fragments, via the AB fragments, to all B fragments without omitting fragments or including any twice before returning to the starting fragment then a full solution is scored. Further details of the algorithm employed will be published elsewhere.

When the search for solutions is complete you can either store the set of solutions, try again at a different error value, or exit. Mirror image solutions are, generally, not stored, and only one of the redundant solutions due to unfixed sites is retained. Unfixed sites are those where two or more 'A' fragments lie totally within a 'B' fragment, such that there are optional positions for the internal A site(s), e.g.:

Fig 1.



*Unfixed site.
(Note that the reverse, or mirror image, of 1 is not the same as 2).

Unfixed sites are very common in double-digest solutions and their presence is one of the reasons for storage of simple double-digest solutions in a temporary file so that more rigorous procedures can be used subsequently in developing a definitive map. Where there are two adjacent unfixed sites it may be noted that in temporary solutions on file these are generally ordered in decreasing order of size to the right, e.g.

Fig. 2

After storage of a set of solutions these can be seen recorded in the temporary area of the main management file, and can be retrieved and viewed at any time using option 2, or selectively deleted using option 5. Their main functions, however, is to act as data sets for option 8.

You may find it instructive to solve a moderately complex double-digest problem, with say, 5 to 8 double-digest fragments, by hand and then compare it to the solution(s) obtained by the program.

Option 8. The purpose of the three enzyme consensus analysis is threefold:
1 to resolve which of the several partial digest solutions for an enzyme pair in the temporary file is the correct one,
2 to resolve unfixed sites, and
3 to provide a second level of checking of site positions so as to make the final maps more accurate and robust.

The result of the analysis is to move temporary mapping data into the final map file. For this the results of all three sets of data giving the relative map positions for three different enzymes are required. A versatile feature of the analysis is that these can be distributed in a number of ways for three enzymes A, B, and C:

1) All three sets of double digest data are present in the temporary file, with the final map file empty. The management file line might look like this:

Test DNA AB2AC1BC4 ----------

2) Two sets of double digest data are present in the temporary file, and the relative map positions of the third pair (BC) are already available from data in the final map file:

Test DNA AB1AC3 ---------- BCDE etc

in this case B and C have already been mapped (together with other enzymes) and the aim of the exercise is to get A fully mapped

3) Three sets of double data are present in the temporary file, with one of these (B in the example below) already mapped in the final file:

Test DNA AB1AC1BC2 ---------- CBED etc

In each case the rigour of the analysis is based on ensuring that an enzyme does not get mapped to the final file until it has been mapped relative to two other enzymes, with all the sites compatible within given error limits.

Approach 1) above is how all maps will start, and by using the TEST 1 data included with the program you can follow the logical progression of the process:

DNA	R.E. pairs in temporary file	R.E.s in final file
1 TEST1	EB2Eg2Bg2HX2EH3EX4SX3HS2Sc1Xc2------	

```
2 T2          ------------------------------------ PSgHXEB
3 T3          ------------------------------------ EBHXS
4 T4          ------------------------------------ BEXSH
5 T5          ------------------------------------ XHBSE
```

TEST 1 is chosen, followed by entering the numbers 1, 2, 3 after the restriction enzyme palette is displayed to choose the enzymes E, B. and g. An error of 0% may then be entered in this case since TEST 1 contains exact dummy data. Normally a good option will be to enter 99, which flags a progressive search with error values increasing in steps of 1% (up to a limit of 10%) until a solution is found. 2% is typically a good value to start with. The program will then first examine the partial maps to select a good order for analysing the map, keeping those with most unfixed sites for later matching. In the example it selects the first (A1) of the two Bg data sets and plots their relative sites as maps 1 and 2 (these figures are depicted on the left of the screen), in mauve on a colour monitor. It then attempts to align the B sites from the first of the two EB data sets with those from the Bg data set by a series of rotation and reversals of the data sets together with tests of each of the possible different positions of identified unfixed sites (2 possibilities for 1 unfixed site, 6 possibilities for 2 adjacent unfixed sites). Sites are accepted as correctly aligned if they are no further apart than E% of the length of the DNA, where E is the current error level set. If alignment of B is successful, the g sites from the Bg map and the Eg map will be tested in a similar way (plotted in red). If these align then the E sites from EB and Eg maps will be tested for alignment (plotted in CYAN), although by now there are more constraints on the manipulation of the sites. There are many unfixed sites in these practice data sets and careful observation of the screen will show the sites flicking from one position to another as the various options are tested. If no alignment of an enzyme is found then data from the next map is tried and if all the trials fail then the screen displays "no fit" or tries all over again after a 1% increment in the error level, if that option was chosen initially. If a complete match is found with all 3 pairs of enzymes then you are given the option of storing the results in the final map file or continuing the search. With real data it is wise to opt for the latter since another solution may be found which is significantly better. The relative merits of such solutions can be assessed by observation of the maps on the screen to see how well the sites align and by noting the values for the mean error, which is simply the mean of the errors at all aligned sites, and the maximum error, which records the largest error found. Continuation of the search is especially recommended if you set a large value (e.g. 5%) for the error at the onset of the search. It is easy to repeat the analysis and stop at the optimum error value once one has a feel for the results of a particular search.

It should be noted that the error concept is somewhat different here than in option 7. Here the error is in the site alignment, and is measured as a % of overall length of the DNA, whereas in option 7 it is measured as a % of the individual fragment length. A good set of data in the three enzyme consensus analysis usually resolves at less than 5% error. Solutions between 5% and 10% should be viewed with suspicion and require thorough checking of the original double-digest data.

Having decided to store a solution for E, B, and g of TEST 1, the site

positions in all 6 maps are displayed for checking and you are asked whether you wish to delete the temporary mapping data which has been used to create the successful map. For real data giving a satisfactory solution it is wise to do this or the data file will become unnecessarily cluttered. While you are practising with the test data, however, or if when analysing real data you are not completely happy with the solution obtained, leave them for possible later use, perhaps in a different combination.

You will now be returned to the main menu and if you call up option 1 you will see that EBg is displayed under the heading of "R.E.s in final file". The map can be displayed in full with option 3.

The test data set can now be used to practise the other approaches to moving temporary data to the final file. The next three sets of double digest data are HX2, EH3, and EX3. These include all 3 possible pairwise combinations of H, X, and E and so are appropriate for a three way analysis, but there is a slight difference in that one of the enzymes, E, is now already mapped to the final file. Perform a three way consensus analysis as above and store the result. The program uses the fully mapped E as a reference for aligning and integrating the new results for X and H with the fully mapped E, B, and g. This introduces an important restriction when using only one reference enzyme to help get two new ones (H and X in this case) into the final map. The reference enzyme must have at least 3 cutting sites, and these must be unevenly spaced, otherwise the new enzyme sites may be aligned or oriented incorrectly. If these conditions are not met a notice to that effect will come on the screen and the results will not be stored.

When H and X are successfully stored the final map listing will look like this : BgEHX

Examination of the next two sets of temporary file data, SX3 and HS2 illustrates use of the third, and in fact the most commonly used, approach to entry of temporary data to the final file. Both X and H have already been mapped to the final file and so the mapped positions of these enzymes in the final file are used instead of the HX2 data in the temporary file (which anyway would in a regular analysis have been deleted by now). So only two sets of temporary file data are used, and the purpose is simply to get the one enzyme, S, into the final file. When asked for the 3 enzymes required, S, X, and H (in any order) are entered in the usual way and the procedure is then the same as above, although the display will note that you are using two reference enzymes from the permanent file. The final enzyme in the Test example is 'c' and this is treated exactly as for S. All the enzymes should now be mapped to the final file and it is a useful experience to delete the whole map file (option 6) and try the consensus analysis in a different order so as to explore what is possible with this data set. See what would happen if you first enter E, B, and g, and next try S, X, and H. The 3 enzyme consensus analysis for the latter will work OK but will not store, because there is no reference enzyme to show how the SXH map relates to the EBg map.

As a general point to re-emphasize : if the data is good, the consensus analysis will find the correct answer, and only rarely will it find two different answers with significant ambiguity. If no solution is found, then the original fragment size data is likely to be erroneous (e.g. have you missed a small fragment?) despite mapping successfully to the temporary file.

Option 9. This performs various analyses of final maps and calls up a short menu of three features:

Analyses of Final Maps

1. Compare individual enzyme site alignments
2. Find phylogenetically informative sites
3. Measure pairwise sequence divergence

Feature one provides another way of looking at enzyme sites in the final map file when there is data on a number of related DNAs. First choose a set of maps, then a restriction enzyme. The sites will be plotted one under the other. This is useful in finding alignments or orientation in new maps and in identifying sites which may be misplaced. Several enzymes can be chosen for superimposition on the same screen and can be useful on a colour monitor, but tends to become rapidly cluttered in monochrome.

Feature two finds phylogenetically informative restriction sites in a set of maps when these are to be used for construction of a phylogeny by a cladistic approach. At least four maps are required since a constraint of a cladistic analysis is that it identifies phylogenetically informative sites only if sites are shared by at least 2 and not more than n-2 taxa (where n is the number of taxa). You then give the error (% of total DNA length) within which you will accept that sites in different maps align. The process then draws up a list of characters. Single sites with no alignments are treated as autapomorphies (a new character evolved in a terminal lineage in the cladist's parlance). Shared sites in no more than n-2 taxa are informative sites, and sites shared by all or all except one of the taxa are assumed to be symplesiomorphies (shared ancestral characters). The assumptions as to autapomorphies and symplesiomorphies may not always be correct but are more likely than the opposite and anyway are not used in the subsequent analyses. A table of phylogenetically informative site positions is drawn up as well as a table of informative character states, where 1 indicates the presence of a shared site, zero the absence of a shared site. You can then, if you wish, give the character set a file name and choose a format appropriate for one of a number of the most popular phylogenetic analysis programs currently available. Do not put an extension on your file name; the program will do that for you with an extension appropriate for your output file choice e.g. choose XX and the file will be XX.PB for PAUP (branch and bound) or XX.HE for Hennig 86 (implicit enumeration). This procedure enable maps to be codified and analysed by phylogenetic analysis programs extremely quickly, a point which will be appreciated by those who have ever tried to do it by hand.

Feature three is used when constructing phylogenies by distance based methods. This requires specifying the lengths of the restriction cutting sequence - usually 6, 5, or 4, or an averaged value if you are unwise enough to complicate matters by using a mixture. In a similar way to feature two you then enter a value for the error in site matching which you consider appropriate, give a file name (no extension), and choose formats from a listing on screen appropriate for your choice of some available phylogenetic construction programs, thus avoiding tedious adjustments to the data file before they can be used, although this will of course, still be necessary if you want to explore more of the options available in these various programs.

Option 10. Toggle display/print modes:
This enables you to print output where this is appropriate. A plotting option for an HP plotter will also be available for the next version.

Option 11. Exit program
It is wise to end a session with this option since it closes all the file buffers and ensures that all output files are fully written to the disc.

Strategy for effective mapping:

Assuming that you are intending to map with a relatively large number of restriction enzymes the best way to proceed is as follows:

1. Perform single enzyme digests on your DNA with all the restriction enzymes which you intend to use and note the number of cuts given by each.

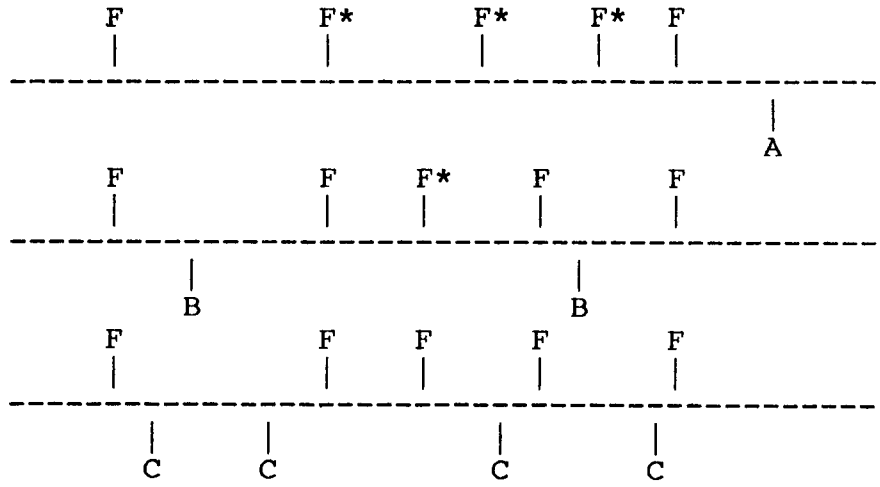
2. Begin your double-digests with three pairwise combinations of three enzymes choosing those which cut the least number of times. Size the fragments by reference to a good set of molecular size markers, with the mobilities of these drawn out with extreme care on preferably semi-log graph paper. The sizes of your restriction fragments are determined by reference to the best line you can draw through your marker points with a flexicurve. There are computer programs which will attempt to do this for you but I am not yet convinced that any but the most sophisticated of these will do better than an intelligent eye and fine pencil (and never use a program which attempts some simplistic linear transformation of your standard curve). Enter the double digest values using option 7 until all three pairs of double-digests are in the temporary file, which will then look something like this : AB2BC1AC2---. Since all three possible pairwise combinations have now been entered you can choose option 8 and if your data is good it will now be transferred to the final map file.

3. Do not now perform digests with three different enzymes. For new enzymes to be aligned with the growing final map you need to take new enzymes one at a time, preferably, and perform two sets of double digests with each of two of the enzymes already mapped to the final file. On successful solution and storage of these temporary maps the temporary file will look something like this : DB1AD3 ----, or CD2DB2---. The three way consensus will work very well on this with the "missing" AB pair (or CB for the second example) being taken from the final map file, and 'D' will then be stored together with A, B and C in the final file. An alternative approach, which can occasionally save some time, is to perform three pairwise double digests in which only one of the three enzymes is already mapped to the final file, e.g. DE1AE1AD3 ---------- CBA. The consensus analysis for D, E, and A will now use the fully mapped A sites as a reference for integrating D and E into the final file, but, as explained above, can only do this if A cuts at least three times.

4. As the map builds up you can work more effectively on those enzymes cutting many times. The art is to use reference (final map) enzymes which are well spaced and do not cut too often, so as to minimise the number of double digest fragments, yet cut sufficiently often so that you minimise the number of unfixed, or 'hanging' sites,

e.g. suppose you were to arrive at these temporary maps using option seven, referencing a new enzyme F against finally mapped A, B, and C :

Fig 4



The best result is b). a) has too many unfixed sites: two adjacent unfixed sites is all that the three enzyme consensus algorithms can handle whereas here there are three c) has 9 fragments (in a circular DNA) which is getting too near the limit for accurate resolution without an unacceptable number of temporary file alternative solutions. More than about eight or nine total fragments in a double digest are beyond the means of any combinatorial algorithm to successfully handle unless it is working on perfectly accurate data, which of couse is not the case in the real world. Hence the need to accumulate early on in the final file a set of well spaced reference enzyme sites with, preferably, one, two, or three sites only for each enzyme. Paradoxically, when mapping a new enzyme with, say, only a single site, double digestion with reference enzymes cutting three or four times may be best since the smaller fragments can be sized more accurately and hence define the new site position most accurately.

5. Enzymes which cut more often than six times may not be resolvable by the program, but are however, often surprisingly easy to map by hand if there a suitable number of reference enzymes already mapped which cut either once or at most twice, and are well spaced. Double digest with the test enzyme and three or four of the reference enzymes. On a print out or plot of the reference map it becomes easy to place the fragments which have been cut, and even those not cut are often easy to place with little ambiguity if they are the only candidate to fill a gap. Multiple small fragments can be a problem and may require use of approaches such as partial digestion mapping, but these more laborious methods can nearly always be avoided by the above approach. When the map has been drawn up to your satisfaction the site positions are simply edited into the final file using option 6.

6. As the maps progress, alignment with respect to other sites, or reorientation in the reverse order of sites, can be evaluated using option 3 or 9, and effected with option 6. Use of enzymes with sites known to be invariant in the DNA set you are using are very useful in this regard, e.g. the Sac II sites in vertebrate mitochondrial DNA. Choose one site and align all the maps on this, but avoid the tendency to edit in other sites in related DNAs if the fragment patterns look the same since this introduces an element of subjectivity which the program is designed to avoid. This can only be allowable if the maps are extremely similar with, say, greater than 90% of sites shared.

7. The solution to the problem of visualisation of restriction fragments is dependent on the quantity of DNA available. There are basically three methods: staining with dyes, end labelling, or southern blotting and hybridisation. End-labelling is a simple and robust procedure and has the great advantage of labelling all fragments in a way independent of molecular size. By preincubating with Klenow fragment of DNA polymerase I in the absence of deoxy-nucleotide triphosphates for 10 minutes before addition of the latter (with one or more labelled) any type of end, 5′ or 3′ overhang, or blunt end, can be readily labelled since the slow 3′ to 5′ exonuclease activity has time to expose an adequate stretch of template for subsequent filling in. End labelling requires a reasonably well purified DNA otherwise interference from e.g. satellite DNAs can be a problem. For further details on this and related problems the reader is referred to the various excellent protocol manuals such as Molecular Cloning : A Laboratory Manual. Eds. Sambrook, Fritsch and Maniatis, or Current Protocols in Molecular Biology. Eds. Ausubel et al.

8. Linear DNA. The program is designed primarily for mapping of circular DNA, however linear DNA can be readily mapped if it is assumed that each enzyme has a cutting site at the end of the linear molecule. A fragment of size zero will have to be included in each set of double-digest fragments to allow for this. When the map is complete (and not before) this cluster of sites at the end of the DNA can be edited out. The main disadvantage is the tendency for a rather larger number of temporary solutions to be found than for circular DNA owing to multiple possibilities of placing the zero sized fragment (a problem common to all cases where there is a very small fragment). Many of these can in practice be edited out, if desired, using option 5 after studying the temporary maps with option 2.

Availability and Distribution

The program will be distributed as an .EXE file, together with data files containing suitable test material for practice, to anyone on request, on either 3.5 inch double-sided double density microdiscs formatted to 1.44 MB or on 5.25 inch double-sided double-density or high capacity mini-flexible discs formatted to 340 KB or 1.2 MB as desired. The program runs on any IBM compatible computer but performs best with a colour monitor and EGA or VGA graphics. Source code will be supplied on request when the program has been accepted for publication. No change will be made for its use other than a nominal $25 for distribution costs (or send a formatted disc) and the .EXE file and data files may be freely copied. However, if the user finds it of value in his/her studies a donation of $50-100 (depending on your means) will be much appreciated and will enable me to place you on file whence you will automatically be sent improved updates as they come out.

Eric H. Harley M.D., Ph.D.
Dept. of Chemical Pathology
Medical School
Observatory
7925 Cape
South Africa

Tel (021) 472150 xt 222
Fax (021) 478955

BIBLIOGRAPHY

Anderson, S., Bankier, A.T., Barrell, B.G., De Bruijn, M.H.L., Coulson, A.R., Drouin, J., Eperon, I.C., Nierlich, D.P., Roe, B.A., Sanger, F., Schreider, P.H., Smith, A.J.H., Staden, R. and Young, I.G. (1981).
Nature, (London) 290 : 457 - 465.

Arnason, U., Gullberg, A., and Widegren, B. (1991)
The complete nucleotide sequence of the mitochondrial DNA of the Fin whale *Balaenoptera physalus*, J. Mol. Evol. 33 : 556 - 568.

Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Seidman, J.G., Smith, J.A. and Struhl, K. (Eds.) (1991)
Current protocols in molecular biology, Massachusetts General Hospital, Harvard Medical School.

Barnes, L.G. (1976)
Outline of Eastern North Pacific fossil Cetacean assemblages, Sys. Zool., Vol 25 : 321 - 343.

Barnes, L.G. (1978)
A review of Lophocetus and Liolithax and their relationships to the Delphinoid family Kentriodontidae (Cetacea: Odontoceti), Nat. Hist. Mus. of Los Angeles County Museum, Sc. Bull., Vol. 28 : 1 -35.

Barnes, L.G. (1984)
Whales, Dolphins and Porpoises: Origin and Evolution of the Cetacea. Reprinted from: Mammals. Notes for a short course organised by P.D. Gingerich and C.E. Badgley. T.W. Broadhead, editor. University of Tennessee, Dept. of Geological Sciences, Studies in Geology, 8 : i - iv, 1 - 234, 1984.

Barnes, L.G. (1984, March)
Search for the First Whale, Oceans, 17(2) : 20 - 23.

Barnes, L.G., Domming, D.P. and Ray, C.E. (1985)
Status of Studies on Fossil Marine Mammals, Mar. Mammal Sci. 1(1) : 15 - 53.

Britten, R.J. (1986)
Rates of sequence evolution difer between Taxanomic groups, Science Vol. 231 : 1393 - 1398.

Brown, W.M., George, M. and Wilson, A.C. (1979).
Proc. Natl. Acad. Sci. USA 76 : 1976 - 1971

Brown, W.M., Prager, E.M., Wang, A. and Wilson, A.C. (1982)
Mitochondrial DNA Sequences of Primates: Tempo and mode of evolution, J. of Mol. Evol. 18 : 225 - 239.

Carlson, S.S., Wilson, A.C. and Maxson, R.D. (1978)
Do albumin clocks run on time ?, Science 200 : 1183 - 1185.

Carr, S.M. and Brothers, A.J. and Wilson, A.C. (1987)
Evolutionary Inferences from Restriction Maps of Mitochondrial DNA from a taxa of xenopus frogs. Evolution 41 : 176 - 190

Collins, W. (1979)
Collins Dictionary of the English Language, Collins, London and Glasgow.

Darnell, J., Lodish, H. and Baltimore, D. (1986)
Molecular Cell Biology, Scientific American Books.

Darwin, C. (1859)
On the Origin of Species, Facsimilied, Atheneum, New York.

Easteal, S. (1991)
The Relative Rate of DNA Evolution in Primates. Mol. Biol. Evol. 8(1) : 115 - 127

Eldridge, N. and Cracraft, J. (1980)
Phylogenetic Patterns and the Evolutionary Process, Columbia niversity Press, New York.

Fitsch, W.M. and Margoliash, E. (1989)
Construction of Phylogenetic Trees. Science 155 : 279 - 284.

Fordyce, R.E. (1980)
Whale Evolution and Oligocene Southern Ocean Environments, Palaeogeogr. Palaeoclimat. Palaeoecol., 31 : 319 - 336.

Fraser, F.C. and Purves, P.E. (1960)
Hearing in Cetaceans, Evolution of the accessory air sacs and the structure and function of the outer and middle ear in recent Cetaceans, Bull. Br. Mus. Nat. His. 7(1) : 1 - 140.

Freifelder, D. (1983)
Molecular Biology: A comprehensive introduction to Prokaryotes and Eukaryotes. Boston, Portola Valley, Jones and Bartlett Publishers, Inc.

Gaskin, D.E. (1982)
The Ecology of Whales and Dolphins, Heinemann Educational Books Ltd., London.

Goodman, M. (1981)
Decoding the pattern of protein evolution. Prog. Biophys. Mol. Biol. 37 : 105 - 164.

Gyllensten, U., Wharton, D. and Wilson, A.C. (1985)
J. Hered. 76 : 321 - 324.

Harley, E.H., White, J.S. and Rees, K.R. (1973b)
The identification of different structural classes of nucleic acids by electrophoresis in polyacrylamide gels of different concentrations, Biochem. Biophys. Acta. 299 : 253 - 263.

Harley, E.H. (1988)
DNA Approaches to Molecular Taxonomy, Trans Roy. Soc. S Afr. 46 Part 4.

Harrison, R. and Bryden, M.M. (eds.) (1988)
Whales, Dolphins and Porpoises, Timmins Publishers.

Hennig, W. (1960)
Phylogenetic Systematics, translated by D. Dwight Davis and R. Zangerl, University of Illinois Press, Urbana.

Hewitt, G.M., Johnston, A.W.B. and Young, J.P.W. (eds) (1990)
Molecular Techniques in Taxonomy, published by Springer-Verlag.

Heyning, J.E. (1989)
Comparative Facial anatomy of Beaked Whales (Ziphiidae) and a systematic revision among the Families of Extant Odontoceti, Contr. Sci., No 405 : 2 : 63.

Hillis, D.M. and Moritz, C. (1990)
Molecular Systematics, Sinaner Associates, Inc., Massachusetts.

Hoelzel, A.R. (ed.) (1991)
Genetic Ecology of Whales and Dolphins, Rep. Int. Whal. Commn., (Special Issue No 13), Cambridge.

Jin, L. and Nei, M. (1991)
Relative efficiencies of the Maximum Parsimony and Distance-Matrix methods of phylogeny construction for restriction data, Mol. Biol. Evol. 8(3) : 356 - 365.

Kasuya, T. (1973)
Systematic Consideration of Recent Toothed Whales based on the Morphology of the Tympano-periotic Bond, Scient. Rep. Whales Res. Inst., Tokyo, No 25.

Kasuya, T. and Marsh, H. (1984)
Life history and reproductive biology of the short-finned pilot whale, Globicephala macrohynchus, off the Pacific Coast of Japan. Rep. Int. Whal. Commn. (Special Issue No 6).

Kellog, R. (1931)
Pelagic mammals from the Trembler formation of the Kern River formation, California. Proc. Calif. Acad. Sci. (Ser. 4), Vol. 19 : 217 - 397.

Kimura, M. (1968)
Evolutionary rate at the molecular level. Nature 217 : 624 - 626.

Kimura, M. (1983)
The neutral theory of molecular evolution, Cambridge University Press, Cambridge.

Lake, J.A. (1991)
Tracing origins with molecular sequences: metazoan and eukaryotic beginnings, TIBS 16 : 46 - 50.

Li, W-H. and Graur, D. (1990)
Fundamentals of Molecular Evolution.

Li, W-H. and Tanimura, M. (1987)
The molecular clock runs more slowly in man than in apes and monkeys.
Nature 36 : 93 - 96.

Li, W-H., Tanimura, M. and Sharp, P.M. (1987)
J. Mol. Evol. 25 : 330 - 342.

Li, W-H., Wolfe, K.A., Sourdis, J. and Sharp, P.M. (1987)
Reconstruction of Phylogenetic trees and estimation of divergence times under nonconstant rates of evolution, Quant. Biol. 52 : 847 - 856.

Lowenstein, J.M. (1985)
Marine Mammal Evolution: The Molecular Evidence (abstract). Sixth biennial conference on the biology of marine mammals.

McClelland, M., Hamish, J., Nelson, M. and Patel, Y. (1988)
Nucl. Acids Res. 16 : 364.

Mead, J.G. (1975)
A Fossil Beaked Whale (Cetacea : Ziphiidae) from the Miocene of Kenya. J. Paleont., Vol. 49, No 4 : 745 - 751.

Mead, J.G. (1975)
Anatomy of the External Nasal Passages and Facial Complex in the Delphinidae (Mammalia : Cetacea), Smithson. Contr. Zool., No 207 : 1 - 63.

Mitchell, E. (1970)
Pigmentation pattern evolution in delphinid Cetaceans: an essay on adaptive colouration, Can. J. Zool., Vol. 48 : 717 -740.

Nei, M. (1975)
Molecular population genetics and evolution, North Holland, Amsterdam.

Nei, M. and Li, W. (1979)
Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. U.S.A. 76 : 5269 - 5273.

Perrin, W.F. (1989)
Dolphins, porpoises and whales; an action plan for the conservation of biological diversity 1988 - 1992. IUCN/SSC Cetacean Specialist Group and U.S. National Marine Fisheries Service, NOAA, Second Edition.

Saitou, N. and Nei, M. (1987)
The Neighbor-Joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol., Vol. 4 : 406 - 425.

Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989)
Molecular cloning, A Laboratory Manual, 2nd Edition, Cold Spring Harbor Lab. Press.

Sarich, V.M. and Wilson, A.C. (1973)
Generation time and genomic evolution in primates, Science, 179 : 1144 - 1147.

Schafer, W. (1972)
Ecology and Palaeoecology of Marine Environments, translated by I. Oetel, Oliver and Boyd, Edinburgh.

Schlötterer, C., Amos, B. and Tautz, D. (1991)
Conservation of polymorphic simple sequence loci in Cetacean species, Nature Vol. 354 issue No 6348 : 63 - 65.

Shimura, E. and Numachi, K. (1987)
Genetic variability and differentiation in the toothed whales, Sci. Rep. Whales Res. Inst., No 38, 1987 : 141 - 163.

Southern, S.O., Southern, P.J. and Dizon, A.E. (1988)
Molecular characterisation of a cloned dolphin mitochondrial genome, J. Mol. Evol., 28 : 32 - 42.

Szalay, F.S. (1969b)
Origin and Evolution of function of the Mesonychid condylanth feeding mechanisms, Evolution, 23 : 703 - 720.

Takahata, N. and Tajima, F. (1991)
Sampling errors in phylogeny, Mol. Biol. Evol., Vol. 8, No. 4 : 494 - 501.

Van Valen, L. (1966)
Deltatheridia, a new order of mammals. Bull. Am. Mus. Nat. Hist., 132 : 1 - 126.

Van Valen, L. (1968)
Monophyly of diphyly in the origin of whales. Evolution 22 : 37 - 41.

Walen, K.H. and Madin, S.H. (1965)
Comparative chromosome analyses of the bottle-nosed dolphin (Tursiops truncatus) and the pilot whale (Globicephala scammoni) Amer. Natur. Vol. 99 : 349 - 354.

Watson, J., Hopkins, N., Roberts, J., Steitz, J. and Weiner, A. (1987)
Molecular Biology of the Gene, Vol. 1, The Benjamin/Cummings Publishing Company, Inc., California.

Whitmore, F.C. and Sanders, A.E. (1976)
Review of the Oligocene Cetacea. Syst. Zool., Vol. 25 : 304 - 320

Wilson, A.C., Ochman, H., Prager, E.M. (1987)
Trends Genet. 3 : 241 - 247.

Wu, C-I., and Li, W-H. (1985)
Evidence for higher rates of nucleotide substitution in rodents than in man
Proc. Natl. Acad. Sci. U.S.A. 82 : 1741 - 1745.

Yablokov, A.V., Bel'kovich, V.M. and Borisov, V.I. (1972)
Kity I Del'finy, Moscos, Izd-vo Nauka. Translated in 1974 (titled Whales and Dolphins) by Joint Publications Research Service, Virginia, U.S.A.

Zuckerlkandl, E., and Pauling, L. (1982)
Molecular disease, evolution and genetic heterogeneity; Horizons in Biochemistry Mikasha and B Pullman (eds.), pp. 189 - 225. Academic press, New York.